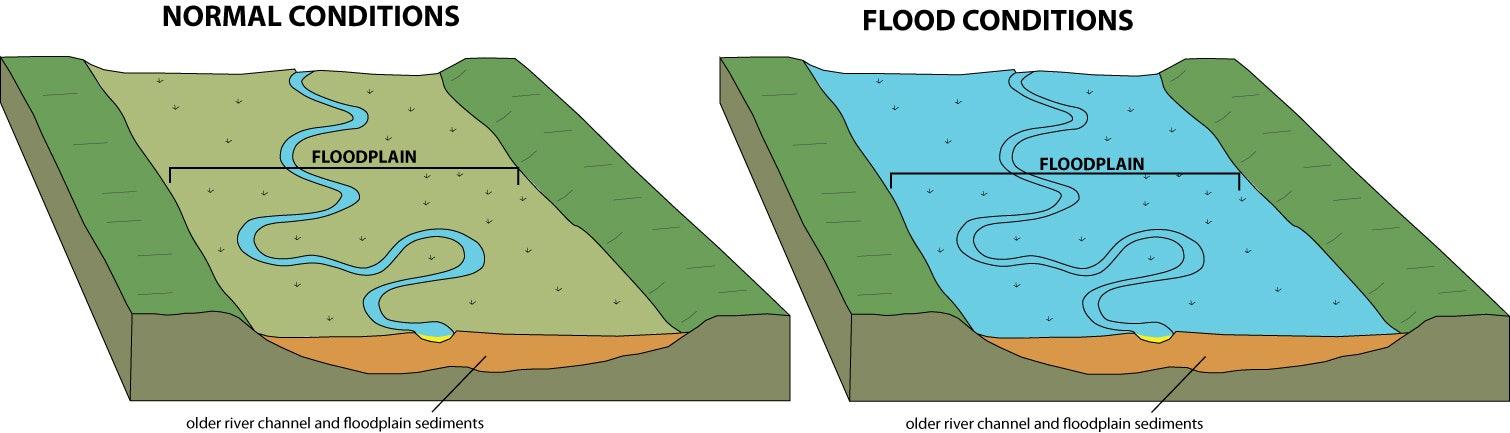
Econometrics Name: Laura Resina (236bf07b)

Dr. Wolf

Final Empirical Project

**Introduction**

You will be asked to specify the independent/explanatory variables and functional form for an equation whose dependent variable is apartment rent (i.e. ). Your goal in this project is to estimate the impact of living in a risky area on apartment rental prices. In particular, you are interested in determining how floodplains impact apartment rental prices. Floodplains are areas where there is at least a 1% chance of a major flood happening every year.



**Figures 1a and 1b: Floodplains**

Floodplains are typically located near major rivers, lakes, or streams and **do not** include areas that are flooded by typhoons or other ocean storms. Floodplains also tend to be in lower elevation areas where it is easy for water to flow to during a major storm.

The cost of living within a floodplain is important to policy makers and economists as it is often used to determine how much money should be spent on flood protection. The greater the risk/cost of living in a floodplain, the more likely the government will try to offset these costs by building flood protection (i.e. dams, levees, flood walls, etc.). In general, it is difficult to calculate this cost as flood risk does not have a market price, though **we expect people living in riskier areas to pay less for their apartments**. Economists have resolved this problem by looking at how apartment rents response to this risk within a hedonic price function (i.e. , where is the coefficient of interest). Differences in apartment rents indirectly allow us to see how much a person is willing to pay to avoid living in a floodplain. **Your goal in this project is to provide the most accurate valuation of that you can using the skills that you have learned from this class. You will begin by estimating an initial model of your own choice (question 3), examining potential problems with this model (questions 2 and 4 - 12), and then estimating a final model that corrects the problems you observed (question 13).**

Before deciding what model, functional form and independent variables to use, it's vital to review the hedonic price literature and to think through the theory behind such models. Such a review is especially important in this case because the model you will be building will be hedonic in nature. A hedonic price model is an equation where you decompose the price of a home or apartment into prices for each of its attributes.

**Literature Review**

Perhaps the most-cited early hedonic price study is that of G. Grether and P. Mieszkowski. Grether and Mieszkowski collected a 7-year data set and built a number of linear models of housing price using different combinations of variables. They included square meters of space, the number of bathrooms, and the number of bedrooms, although the latter turned out to be insignificant. They also included lot size (i.e. how much land the property came with) and the age of the property as variables, specifying a quadratic function for the age variable. Most innovatively, they used several slope dummies in order to capture the interaction effects of various combinations of variables. Peter Linneman estimated a hedonic price model on data from Los Angeles, Chicago, and the entire United States. His goal was to create a model that worked for the two individual cities and then to apply it to the nation to test the hypothesis of a national housing market. Linneman did not include any lot (land) characteristics, nor did he use any interaction variables. His only measures of the size of the living space were the number of bathrooms and the number of non-bathrooms. Except for an age variable, the rest of the independent variables were dummies describing quality characteristics of the property and neighborhood (i.e. whether there was a park nearby, if the house was built on a hill, etc.). Although many of the dummy variables were insignificant, the coefficients of age, number of bedrooms were statistically significant and substantially lowered the model’s residual sum of squares.



K. lhlanfeldt and J. Martinez-Vasquez investigated sample bias using various types of pricing data (i.e. the sale price of a property versus the asking price). Unfortunately, they went on to estimate a regression equation with a large number of independent variables and then dropping all those that had t-scores below one, almost surely introducing bias into their equation since some of the omitted variables could have arguably been relevant/significant in another sample. Raymond Palmquist added some innovative variables to an estimate on a national data set. He included dummy variables indicating whether a property had a parking space, an air conditioning unit, and if the property was considered to be in excellent condition. Perhaps most importantly he found that the age and size of the property were non-linearly related with price.



Finally, when reviewing the flood risk literature, you find that apartment rents and housing prices are **not** lower in floodplains unless there is a recent natural disaster such as a typhoon to remind residents of the risk they face. After a natural disaster, floodplain apartment rents are typically 3% to 10% less than non-floodplain apartment rents (Bin and Landry, 2013; Atreya et al., 2013). **You further find that one of the most powerful typhoons to have ever hit Japan – Typhoon Prapiroon - occurred during June and July of 2018. Typhoon Prapiroon caused significant damage all across the country but was especially damaging to homeowners and renters living in Fukuoka, Hiroshima, and Ehime Prefecture.**

**Model**

Now that you’ve reviewed at least a portion of the literature, it’s time to build your own model. Here is a sample model which relates the rent of an apartment to its square footage and whether it’s in a floodplain.

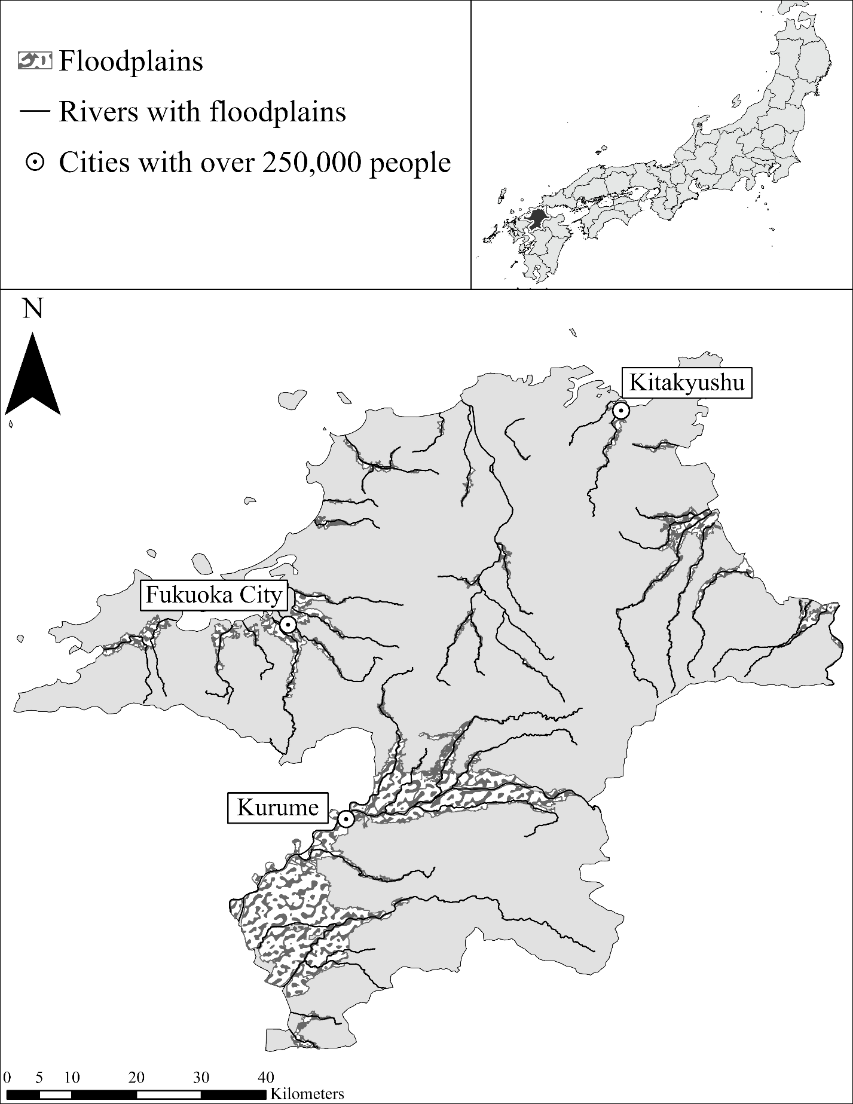
Where

the monthly rent of apartment i. Measured in Japanese Yen and includes the deposit fee.

the size of the ith apartment measured in square meters

an indicator whether apartment i is located in a floodplain. A floodplain is an area where there is a higher chance than normal of a flood occurring.

The above equation was estimated using the same sample of 264,826 apartments that is in your dataset (Final Project Data.Rdata). All of these apartments in the dataset are located in Fukuoka Prefecture (See Figure 1) and were leased between 2015 and 2019.



There are several additional independent variables that could be incorporated into the above model which you have access to in your dataset:

**the year apartment i was built**

**indicator whether apartment i has a living room (i.e. think 3LDK)**

**indicator whether apartment i has an air conditioning unit**

**the number of months apartment i was listed for rent prior to being leased**

**the floor or story the ith apartment is located on**

**indicator whether apartment i has a dining room (i.e. think 3LDK)**

**indicator whether apartment i has bicycle parking nearby**

**the year in which the apartment sold**

**the month in which the apartment sold**

**indicator whether apartment i is a one-room unit (ワンルーム)**

**how many bedrooms there are in apartment i (i.e. think 3LDK)**

**indicator whether the building the apartment is in has an elevator**

**indicator whether the apartment building has security cameras**

**indicator whether the apartment building has fiber internet (high-speed internet)**

**indicator whether the apartment building is made from steel**

**indicator whether the apartment building is made from concrete**

**the number of minutes it takes to walk to the nearest train station**

**distance from apartment i to the nearest river, measured in meters**

**distance from apartment i to the nearest park, measured in meters**

**Neighborhood identification number of apartment i.**

**Questions**

**Note: The answer to each question must be completed on the computer. I will not accept projects that have hand-written answers. You must show all of your work to receive full credit. This includes copying regression/statistical output from R and pasting it below each question.**

1. Read through the list of variables, making sure you understand the theory behind using each variable. Write out the expected signs of the coefficients on , **,** , and ? For each of the listed variables above write out the null and alternative hypothesis statements. Explain your decisions.

:

𝐇𝟎:𝜷>=𝟎 against 𝐇𝐚:𝜷<𝟎 → Here I expect that the closest to the station, the higher the monthly rent. Or, the more minutes it takes to walk to the nearest station, the lower the rent.

:

𝐇𝟎:𝜷=𝟎 against 𝐇𝐚:𝜷 ≠𝟎 → I do not have any expectation in regards to how the distance from a river influences the rent price. Therefore, I use the ≠ symbol to show I have no prior expectation.

:

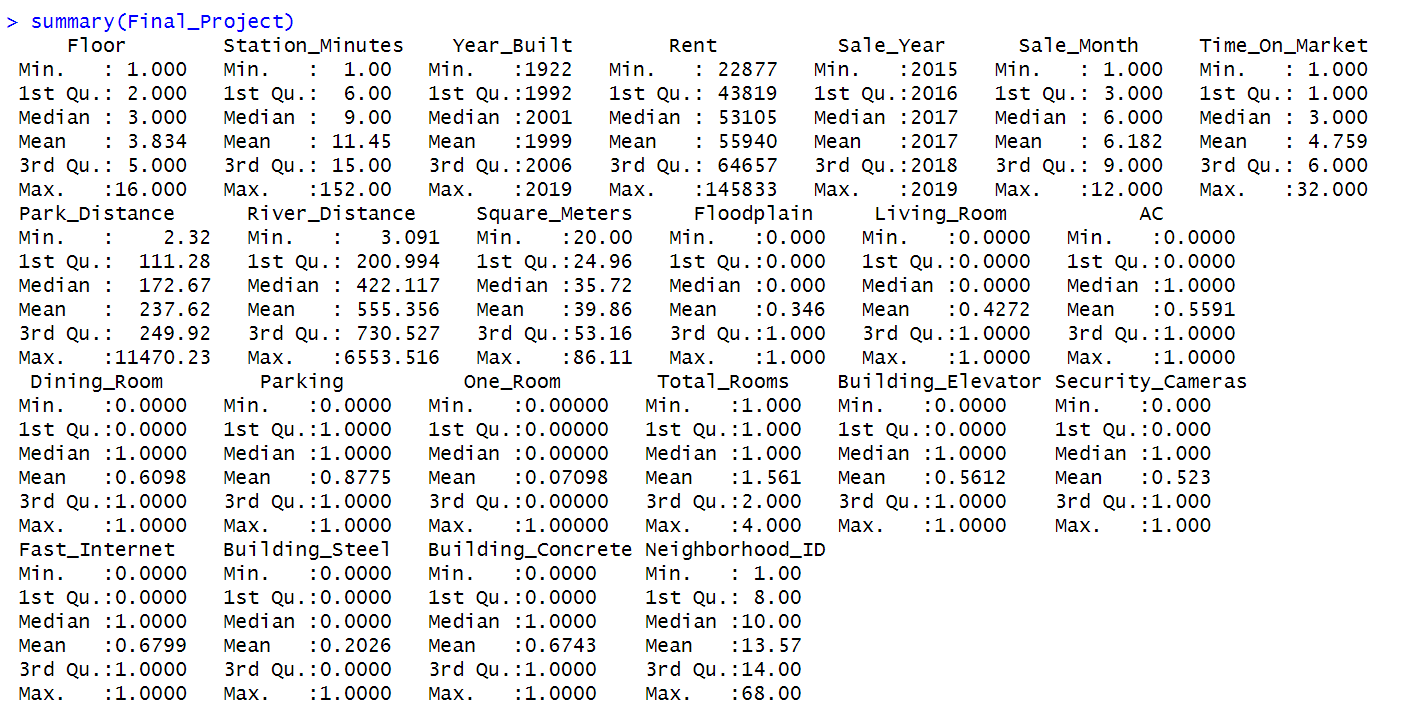
𝐇𝟎:𝜷>=𝟎 against 𝐇𝐚:𝜷<𝟎

I expect that a one room apartment is more likely to have a lower rent price

𝐇𝟎:𝜷𝟓<=𝟎 against 𝐇𝐚:𝜷𝟓>𝟎 → Here I expect that if the building where the apartment is has an elevator, the rent price would be higher.

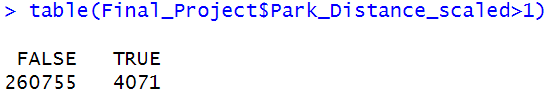
1. Describe the average apartment in your dataset (i.e. what is the average rent, square meters, age, number of bedrooms, etc.). Provide summary statistics (i.e. min, mean, max, etc.) below. Are all the variables measured in acceptable units (i.e. do you need to scale any of your variables)? Do you think any of the apartments are outliers (i.e. have extreme characteristics or have values that do not make sense? **Please make sure to check all of the variables when searching for outliers**)? If you think there are outliers, please remove these observations from the dataset and explain why you did this.

**Firstly, I summarized all the variables available in the dataset. The results are as follows:**





* **Units check:**

**Park distance and river distance could be measured in kilometers, since it would not make sense to interpret how much rent prices increases if increasing by 1 meter either park distance or river distance. During the process of deciding such, I decided to scale even if the percentage of houses with a parking spot at a higher distance than 1000 m (1km) is just 1.53%. .**

**On the other hand, the percentage of houses at a distance from the river above 1 km instead is around 13.41%. I will scale it since it would be easier to interpret the results. A close-up of a computer screen

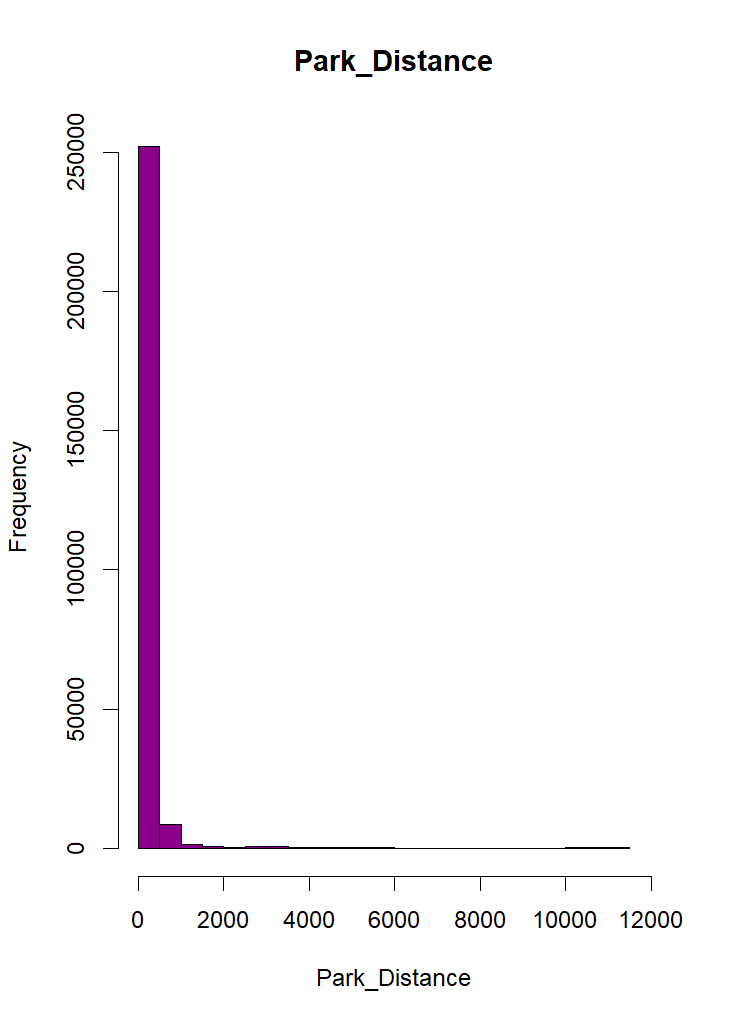
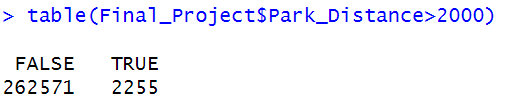
Description automatically generated .**

**For the rest, I do not think a scaling is needed since they are all measured in acceptable units.**

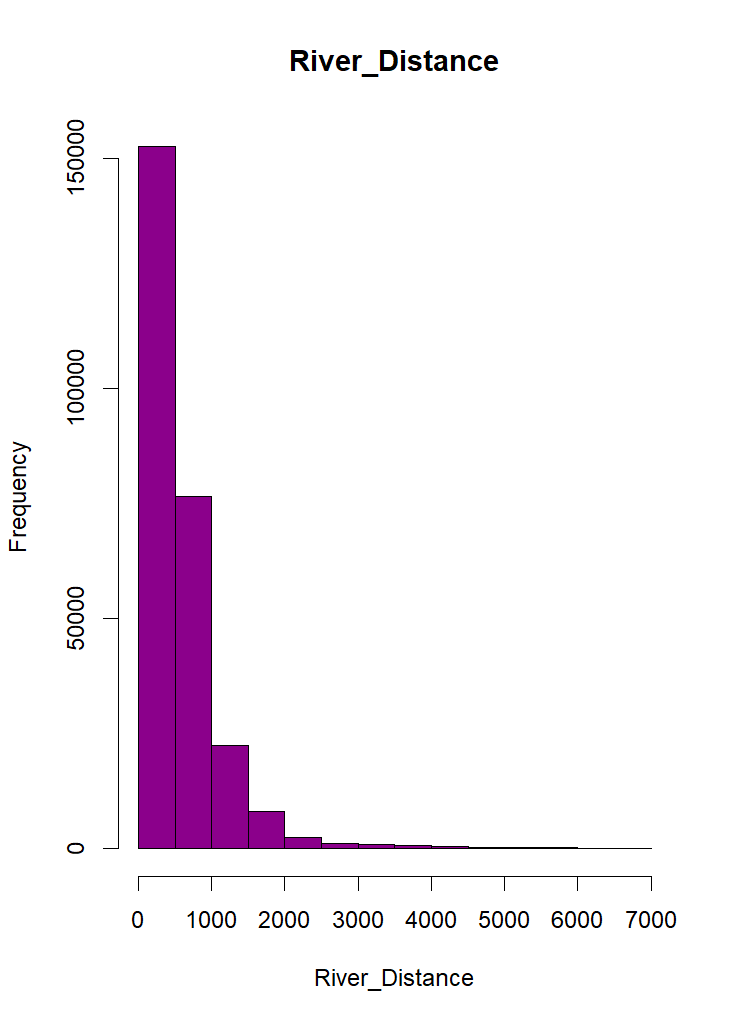
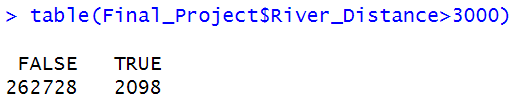
* **In relation to the outliers, I checked all the ranges of every variable, and firstly compared the median with the maximum and minimum value. Then, Assuming a normal distribution, for the variables that had a maximum value very high compared to the mean I gave it a closer look.**

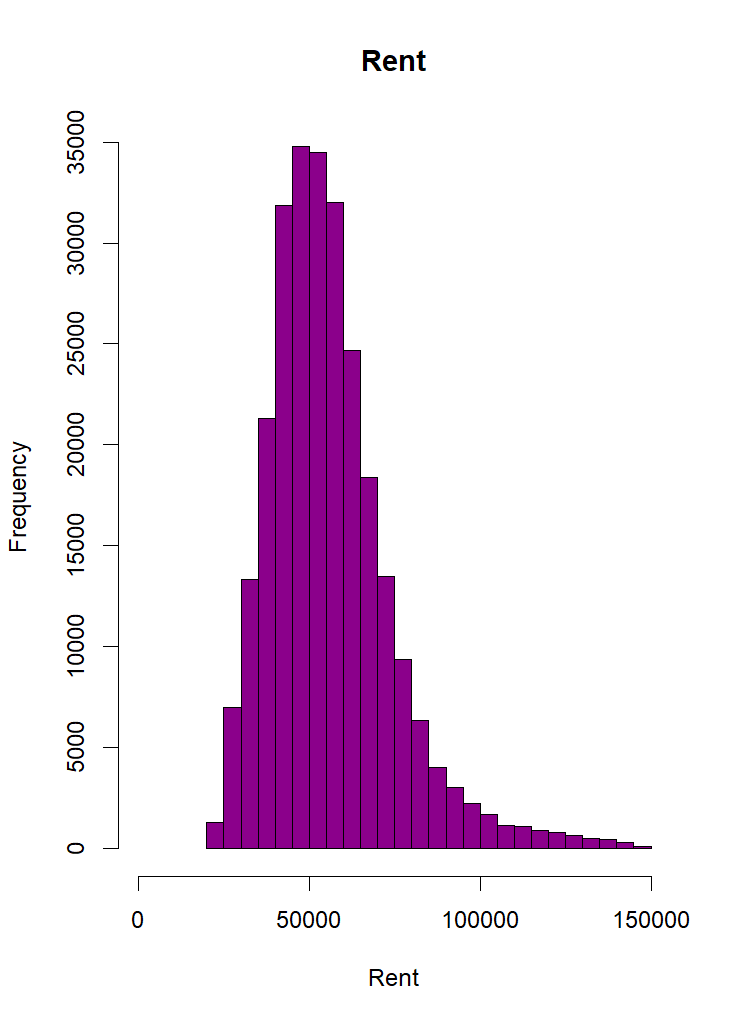
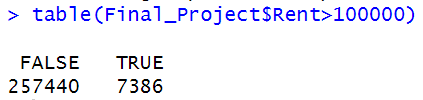
**-Park\_Distance and River\_Distance seem to have very high max values compared to the min and the median.**

**For Park Distance\_scaled I will remove the outliers above 2 km, which represent around just the 0.8 % of the sample and also because it would not seem rational to have a parking lot above 2 km**

**** ****.

**For River Distance\_scaled instead, I will cut the outliers above 3000 meters (3km),since it represent just around the 0.8 % of the sample.**

**** ****

**-Rent max is way higher than the median, likely to have outliers, as also shown from the grap. I will remove the Rent above 110000 which represent around the 1.73% of the sample.** ****

1. Write out your initial regression equation and then estimate it using R. This should be your first guess of your regression model (i.e. it does not need to be perfect. In fact you will try to improve upon this model when completing questions 4 – 12). Copy and paste the output from this model below. Are all your independent variable coefficients significant? Does the magnitude on your coefficients make sense? Does your estimate of flood risk match the predictions found within the literature?

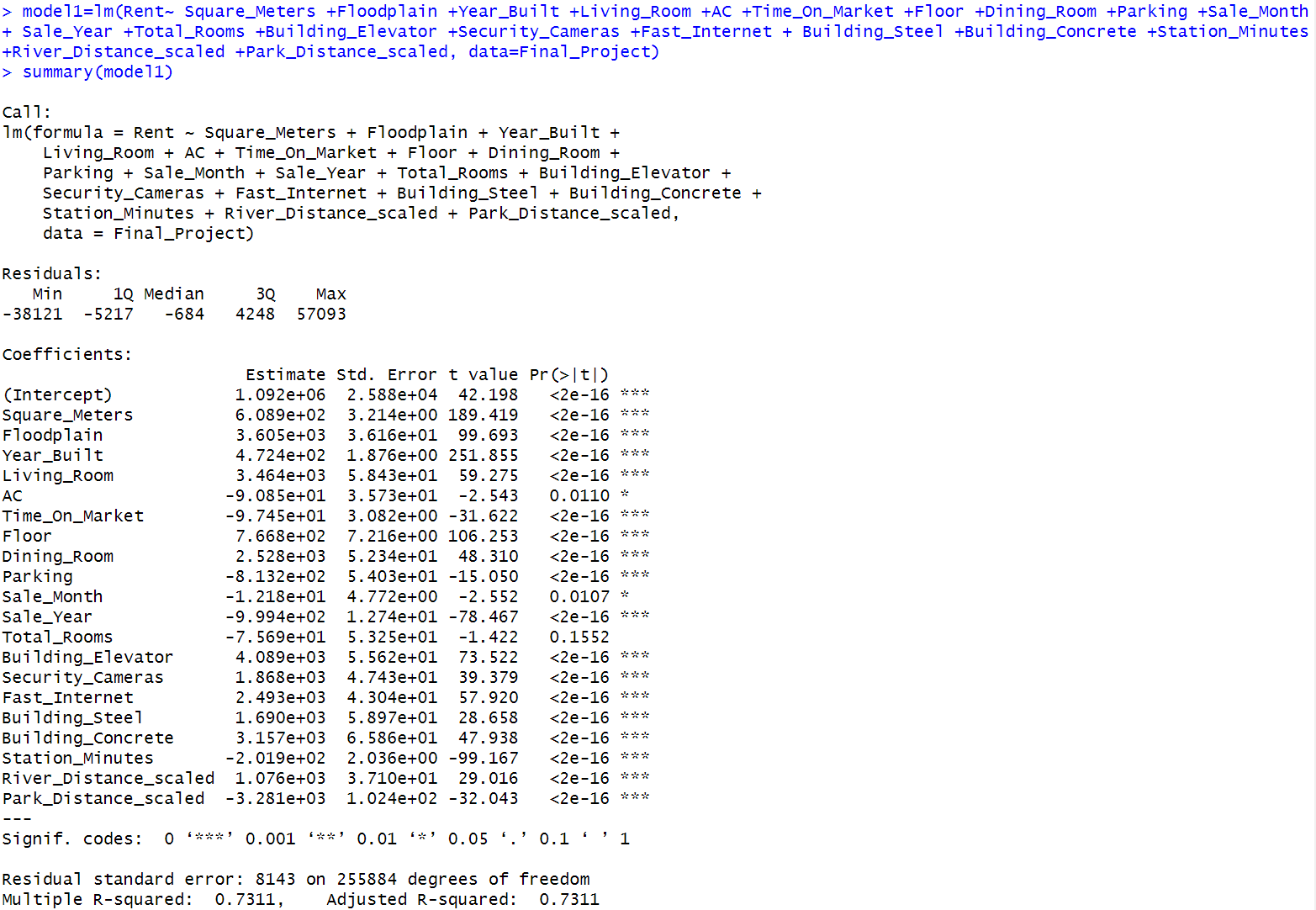
* **In my regression model I am adding as much variables as possible apart from the following:**

**-Neighborhood\_ID (because I think it is not relevant in relation to the rent price),**

**-One\_Room (because I suppose a correlation with Total Rooms, and I could get the information already from the Total\_Rooms, and prevent multicollinearity and consequently type 2 error).**

* **Code:**

**model1=lm(Rent~Square\_Meters+Floodplain+Year\_Built+Living\_Room+AC+Time\_On\_Market+Floor+Dining\_Room+Parking+ Sale\_Month + Sale\_Year+Total\_Rooms+Building\_Elevator+Security\_Cameras+Fast\_Internet+Building\_Steel +Building\_Concrete+Station\_Minutes+River\_Distance\_scaled +Park\_Distance\_scaled, data=Final\_Project)**

****

* **SIGNIFICANCE OF THE COEFFICIENTS:**

**All the coefficients are significant.**

* **INTERPRETATION OF THE COEFFICIENTS:**

1. **Square\_Meters: For each additional square meter of the property, Rent is estimated to increase by approximately 651.9 yen.**
2. **Floodplain : holding all other variables constant, Rent is estimated to increase by 3811 yen when located in the Floodplain compared to the properties that are not located in a floodplain (rationally it does not make sense, I would have expected a negative relationship, but it does reflect the literature as per the apartment in a floodplain does not influence the rent overall, unless there is a recent natural disaster).**
3. **Year\_Built : For each additional year the property was built (assuming other variables are held constant), the Rent is estimated to increase by approximately 488 yen. Which means that the most recent buildings costs more and it resonates.**
4. **Living\_Room : Having a living room in the property (assuming other variables are held constant), is estimated to increases Rent by approximately 3,021 yen.**
5. **AC : Having an air conditioning in the property is associated with an estimated lower Rent of approximately 211.9 yen compared to the one that don’t have it, holding all other factors constant. (it does not resonate, but at the same time having an AC does not seem a relevant factor considering the low impact it would have on rent).**
6. **Time\_On\_Market: For each additional month the apartment stays on the market (holding all other variables constant), the Rent is estimated to decrease by approximately 108.2 yen. Annually it would have an impact on rent of approximately 1298.4 yen, which is quite low. This negative coefficient implies that apartments with longer time on the market may have slightly lower rent values.**
7. **Floor: For each additional floor of the apartment (holding all other variables constant), the Rent is estimated to increase by approximately 930.5 yen. It indicates that apartments on higher floors tend to have higher rent values.**
8. **Dining\_Room: By having a dining room (holding all other variables constant), the Rent is estimated to increase by approximately 1250 yen. Apartments with dining rooms tend to have higher rent values.**
9. **Parking: Having parking facilities in the apartment is associated with an estimated decrease in Rent of approximately 907.6 yen. This negative coefficient indicates that apartments with parking may have slightly lower rent values compared to those without parking (It does not resonate with my expectation of positive sign).**
10. **Sale\_Month: For every one-unit increase in Sale\_Month (assuming it is a numerical variable), the estimated Rent decreases by 18.8 yen, all other factors being equal. This does not make sense because it would mean that in December rent would be lower (it is not a continuous variable).**
11. **Sale\_Year:** **For every one-unit increase in Sale\_Year (assuming it is a numerical variable), the estimated Rent decreases by 1014 yen, all other factors being equal.**
12. **Total\_Rooms: For each additional total room in the apartment (holding all other variables constant), the Rent is estimated to decrease by approximately 513.2 yen. This negative coefficient suggests that apartments with more rooms may have slightly lower rent values. (It does not make sense to have negative coefficients, but at the same time the magnitude is very low and statistically insignificant). That is why I will drop it.**
13. **Building\_Elevator: Having an elevator in the building is associated with an estimated increase in Rent of approximately 4,684 yen. This positive coefficient indicates that apartments in buildings with elevators tend to have higher rent values.**
14. **Security\_Cameras: Having security cameras in the apartment is associated with an estimated increase in Rent of approximately 2073 yen. This positive coefficient suggests that apartments with security cameras tend to have higher rent values.**
15. **Fast\_Internet: Having fast internet access in the apartment is associated with an estimated increase in Rent of approximately 2,597 yen.**
16. **Building\_Steel: Having a steel building construction is associated with an estimated increase in Rent of approximately 1,486 yen. This positive coefficient suggests that apartments in steel buildings tend to have higher rent values.**
17. **Building\_Concrete: Having a concrete building construction is associated with an estimated increase in Rent of approximately 2687 yen.**
18. **Station\_Minutes: By increasing the minutes to reach the station by 1 unit(holding all other variables constant), the Rent is estimated to decrease by approximately 237.5 yen.**

**If I increase the distance from the station by 1 hour, the rent is estimated to decrease of approximately of 237.5 \*60=14250 yen. This negative coefficient suggests that apartments farther from a station may have slightly lower rent.**

1. **River\_Distance\_scaled: For each additional km of distance from the river (holding all other variables constant), the Rent is estimated to increase by approximately 659.8 yen.**
2. **Park\_Distance\_scaled:** **For each additional km of distance from the park (holding all other variables constant), the Rent is estimated to decrease by approximately 1620 yen.**

**This negative coefficient suggests that apartments farther from the parking lot may have lower rent values. It does not seem relevant in magnitude.**

* **Magnitude of Coefficients:**

**The magnitude of the coefficients makes sense. The coefficients for Square\_Meters, Living\_Room, Floor, and Station\_Minutes are relatively large, indicating that these variables have a substantial impact on the Rent. On the other hand, the coefficients for AC and Sale\_Month are smaller, suggesting that these variables have a relatively smaller effect on Rent.**

**Also River\_Distance\_scaled, Park Distance, Fast internet, Time on Market, Total\_Rooms, Parking have relatively low impact.**

**Interpretation of Flood Risk (β̂\_Floodplain):**

* **The t-value for Floodplain is very high, and the p-value is very low, indicating that the coefficient is highly statistically significant at 0.01% and a significant predictor of Rent. According to the results though, properties in the floodplain tend to have higher rent values compared to those that are not in the floodplain. This goes against my expectations but it resonates with the predictions found within the literature where the apartment rents and housing prices are not lower in floodplains unless there is a recent natural disaster such as a typhoon. (Bin and Landry, 2013; Atreya et al., 2013).**

1. Check to see if you have any irrelevant variables. How might this problem affect your conclusions. Which variables seem potentially redundant?

**Firstly, the first thing I looked at was the statistical significance, which according to the results, all the variables seem statistically significant. At the same time I analysed the magnitude of each variable.**

**- Neighborhood\_ID is not relevant in relation to the rent price. (No impact on R-squared by adding it, nor much influence on the Standard errors).**

**- One\_Room is potentially redundant because I could already get the information from the Total\_Rooms.**

**-Total Rooms** **could be dropped since the impact on rent for adding one more room is too low to be relevant. Dropping it does not influence the adjusted R-Squared. Also Squared Meters and Total Room together could be redundant.**

**-Sale\_Month: according to the findings and the magnitude, it does not seem to have a big impact on the price. I will drop it. Plus, it is not categorical.**

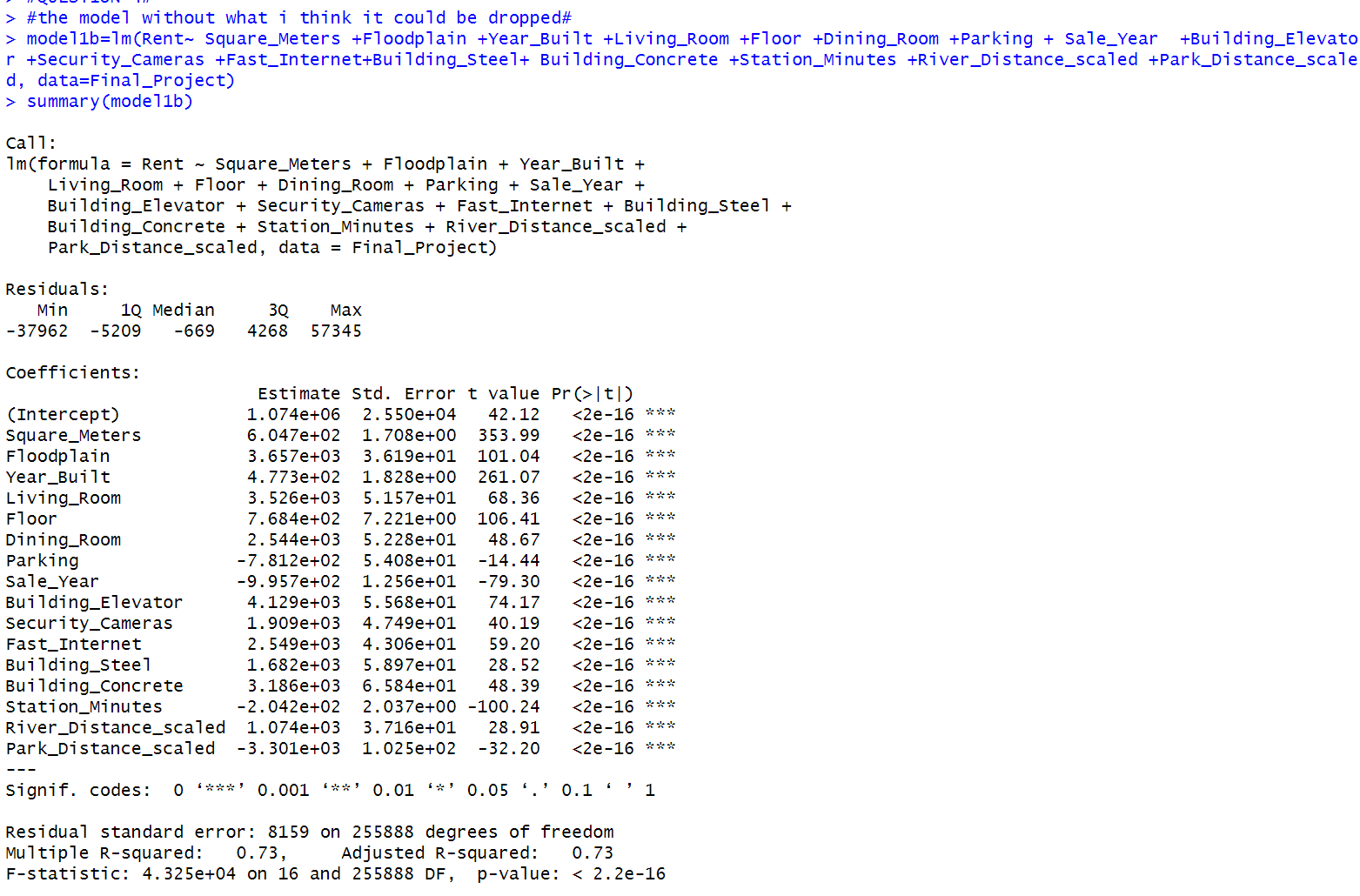
**-AC ,even if statistically significant, is not practically meaningful and has a low impact on rent prices, so it could be dropped. Checking the model without AC, Adjuster R-squared is not impacted. (test AC)**

**-Time on Market: considering the low magnitude of impact on rent it could be dropped. By running a test of dropping it, is barely influencing the model , the SE or the adjusted R-squared.**

**The irrelevant variables could lead to less precise estimate, an increase in SE, a decrease t value, and consequently more likely to commit type 2 error.**

**By adding variables that are not needed or redundant, could lead to the dummy variable trap and make it hard to estimate considering the perfect collinearity that could be caused.**

**TEST: model 1b**

****

1. Are there any new variables you could make from the information provided in the dataset that you’d like to include in the model? Are there any variables you omitted initially but now think you should include? What happens to the regression when you include these variables?

-I added an After Typhoon dummy variable to track the difference in prices before and after the typhoon. By looking at the literature, it actually hit Japan in July, and that is why I used July as the month of reference.

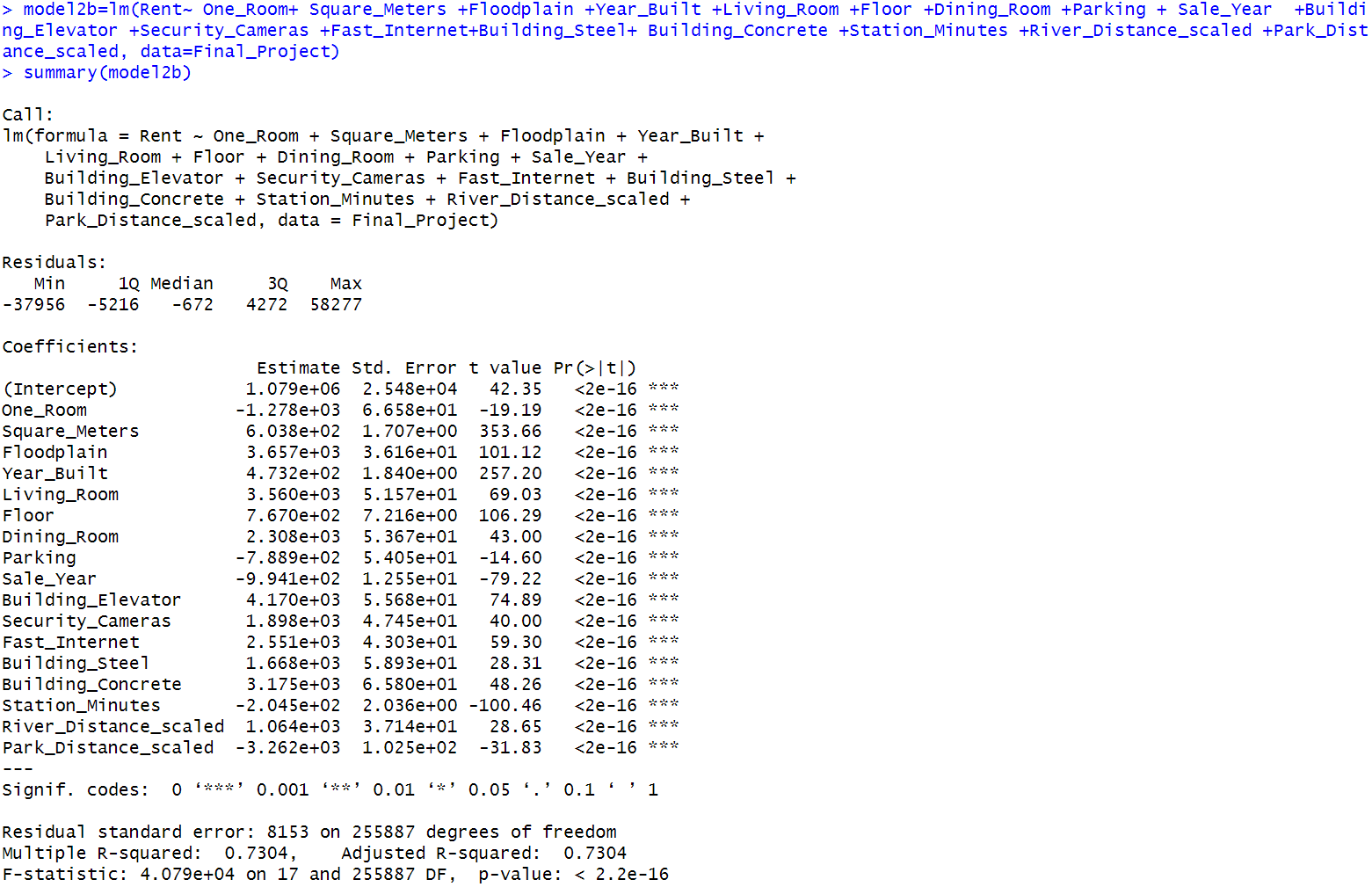
- I am including is the age variable of the building at the moment it was sold (Sale Year- Year Built) variable. The building built in 2019, will have a value of 0 as they are brand new.

* Variables I omitted initially and that I am now including:

By testing for example the inclusion of One\_Room to the model 1b ,

my initial supposition of correlation of one room with total rooms, was incorrect as the data shows: we see a weak negative correlation, which would not lead to multicollinearity if I were to include both variables. As for now and for the irrelevance of the impact total rooms had, I will only include One Room.

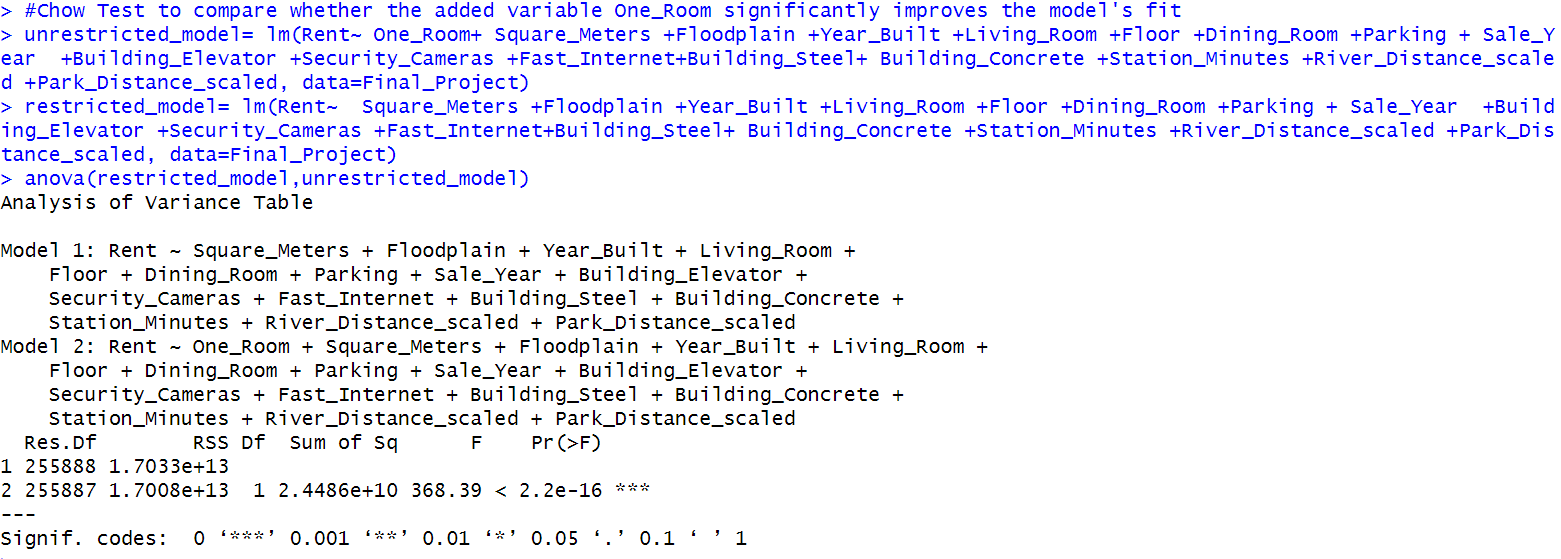




As a result, One\_Room seems to be statistically significant at the 0.1%.

I also performed a chow test, to understand if there is a significant difference between the 2 models. The unrestricted model will be the one with One\_Room added, whilst the restricted won’t have One\_Room.

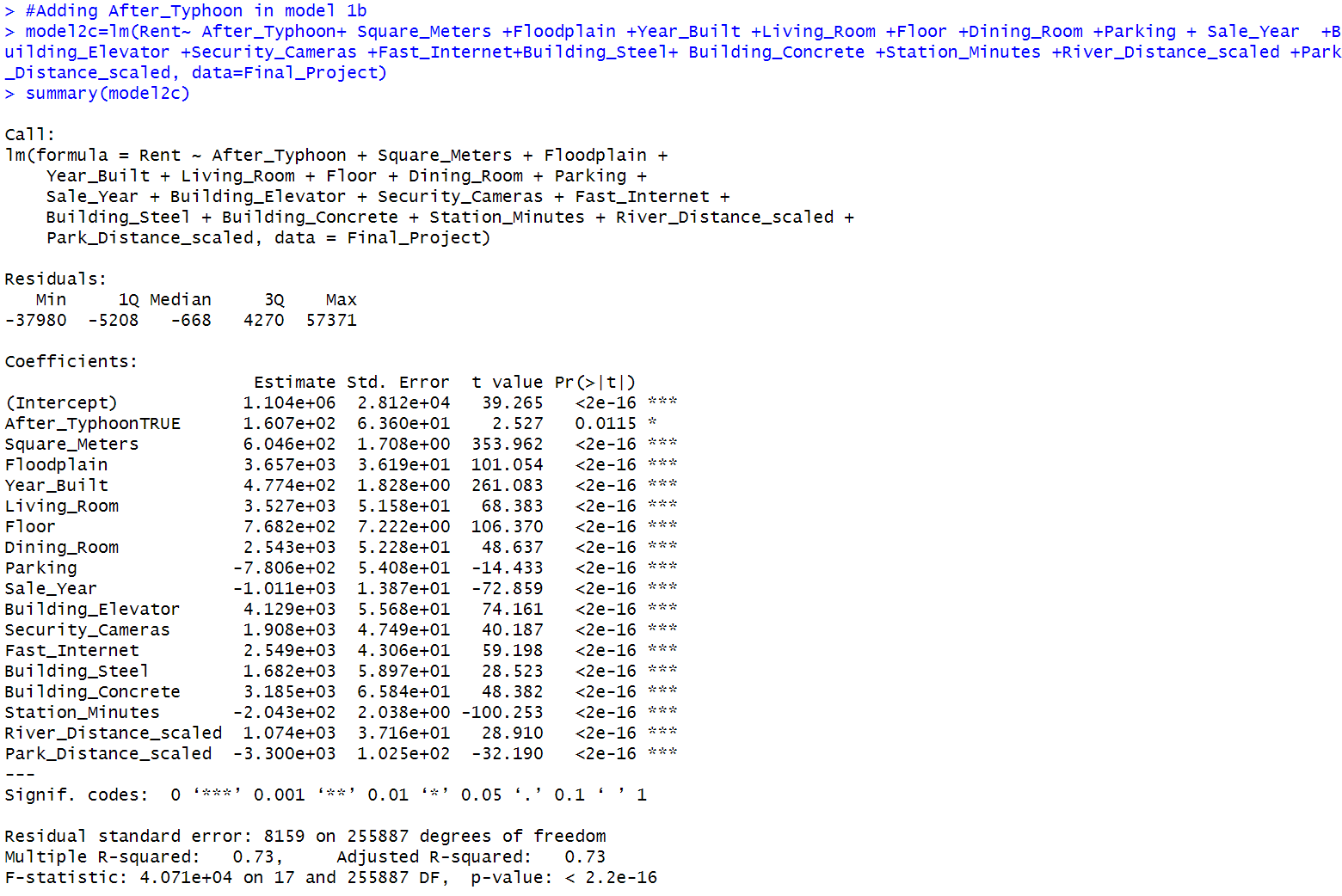
The null hypothesis of the Chow test is that there is no difference between the two models, indicating that the coefficients are stable and there is no structural change. The alternative hypothesis is that there is a significant difference between the two models, suggesting that there is a structural change and that the coefficient of "One\_Room" is significantly different from zero.

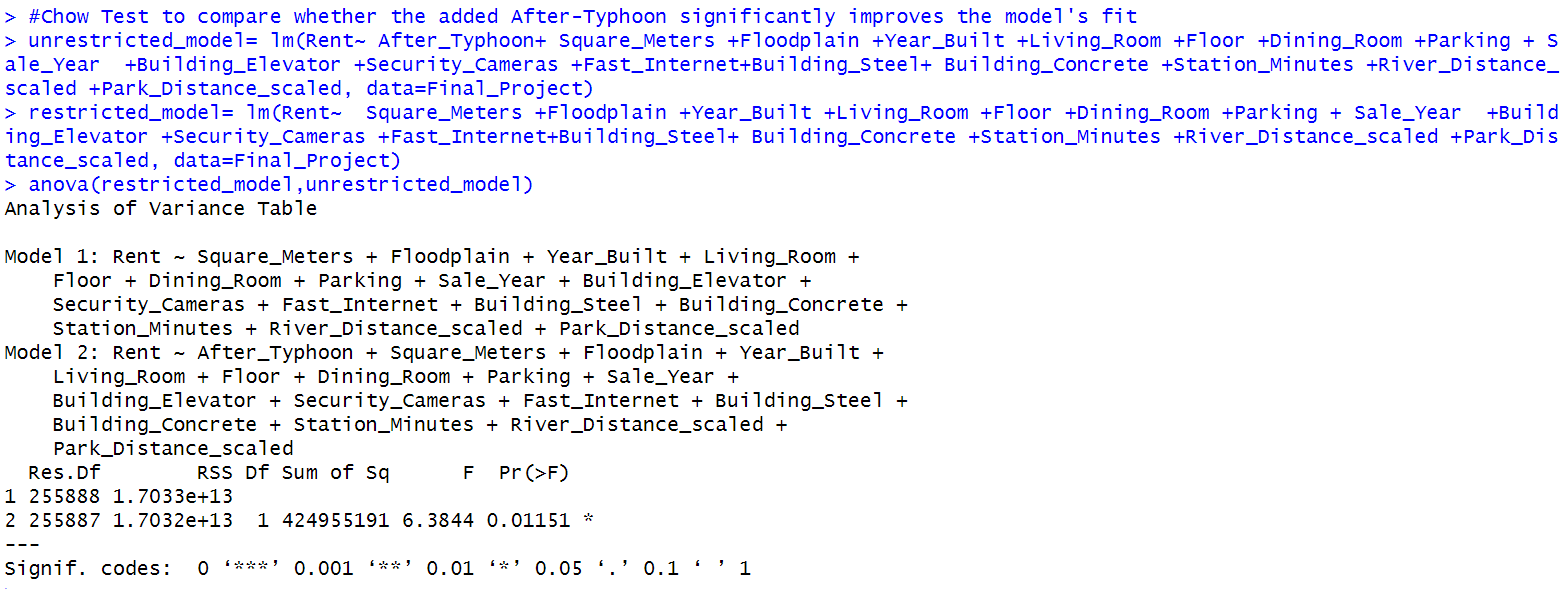


Looking at the results of the p-value, it is very low and statistically significant at the 0.1%, which leads me to reject the null hypothesis and accept the unrestricted model with included the One\_Room.

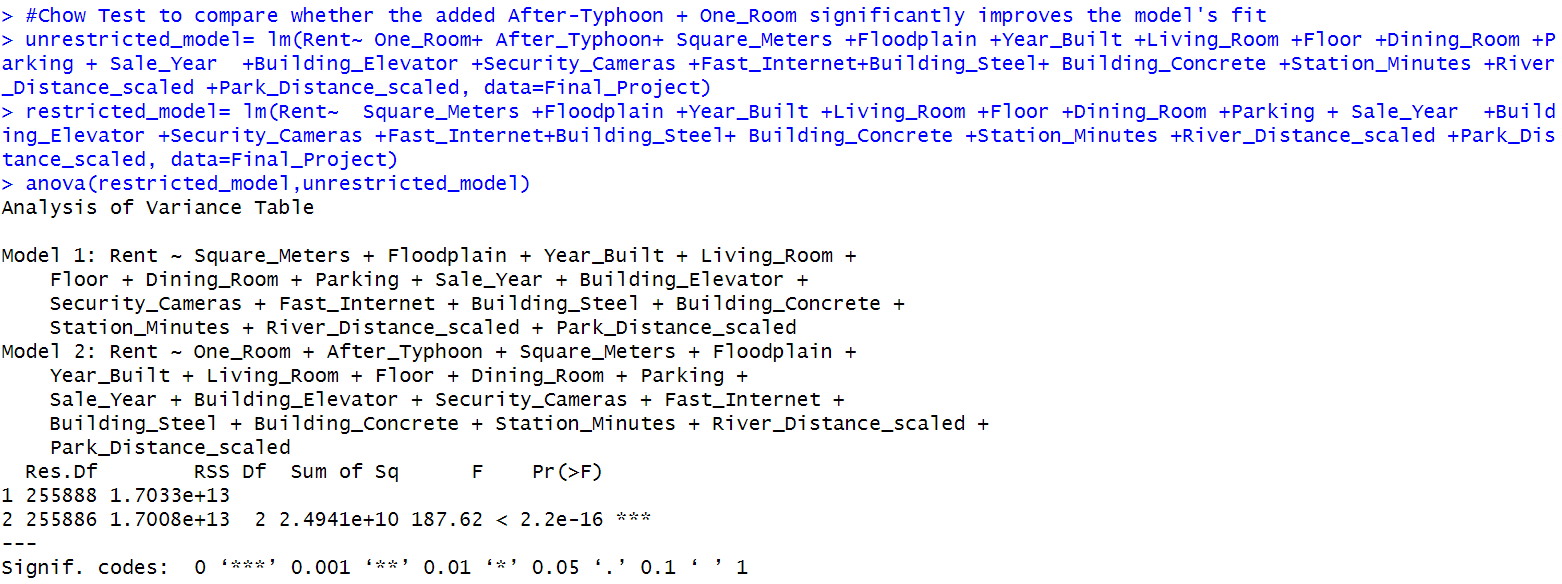
In this case, I realized the importance of One\_Room.

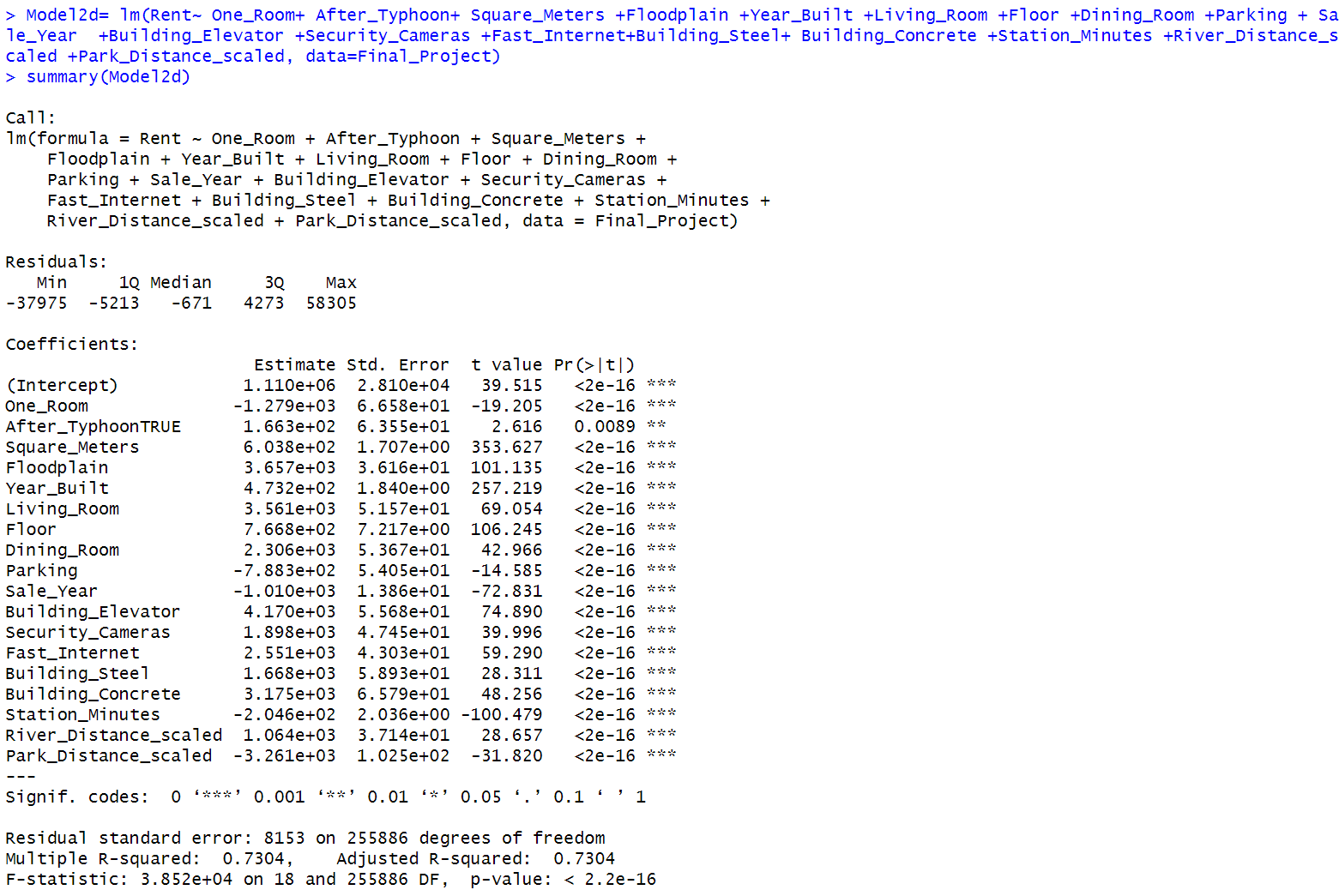
My new model will be the following “model2b”.

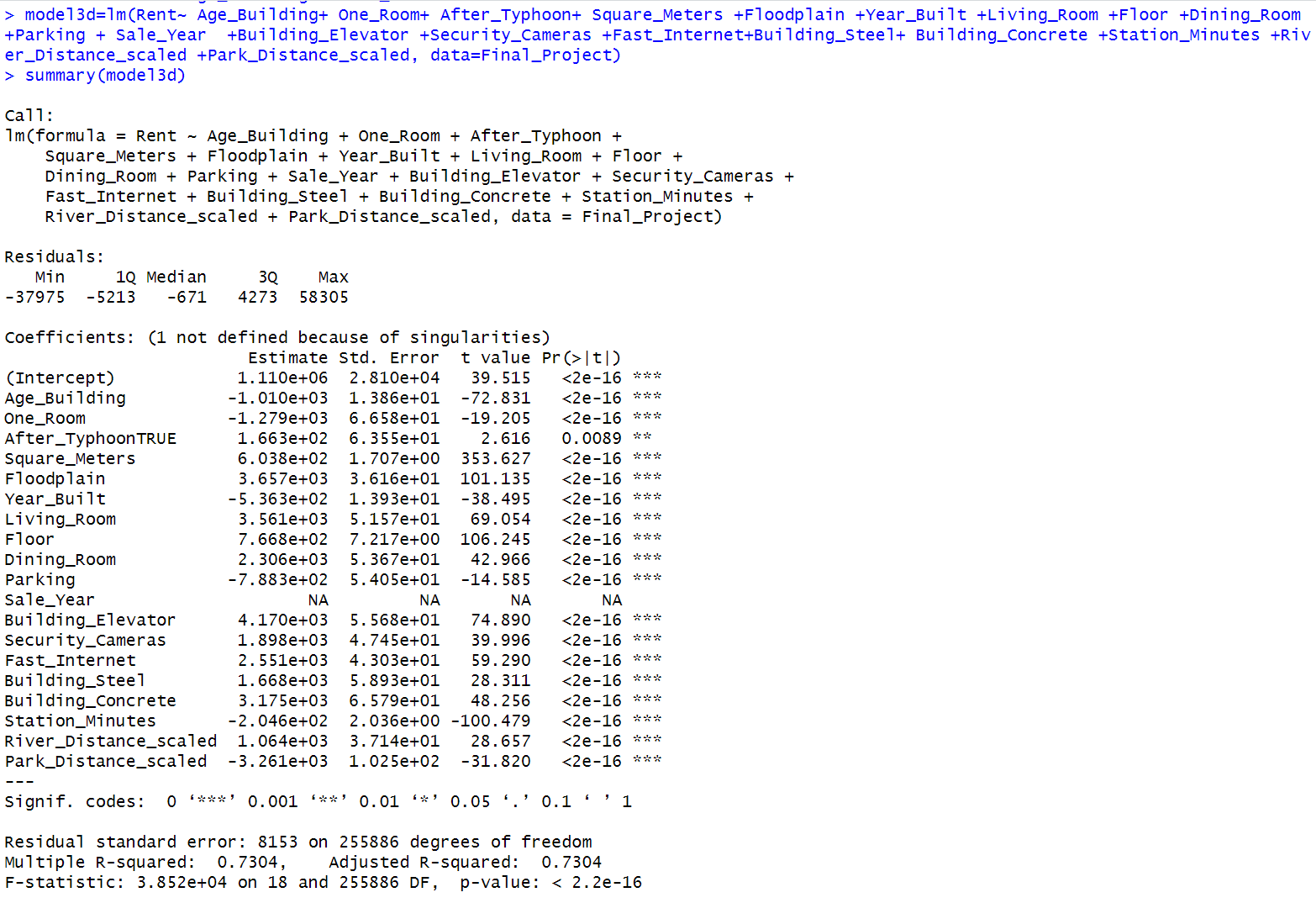
* Analysing the effect of adding After\_Typhoon: it is statistically significant at the 5%. And also looking at the chow test, the p-value is 0.005063, which is less than the conventional significance level of 0.05. Therefore, we reject the null hypothesis in favor of the alternative hypothesis. This means that the variable "After\_Typhoon" does significantly improve the model's fit. 



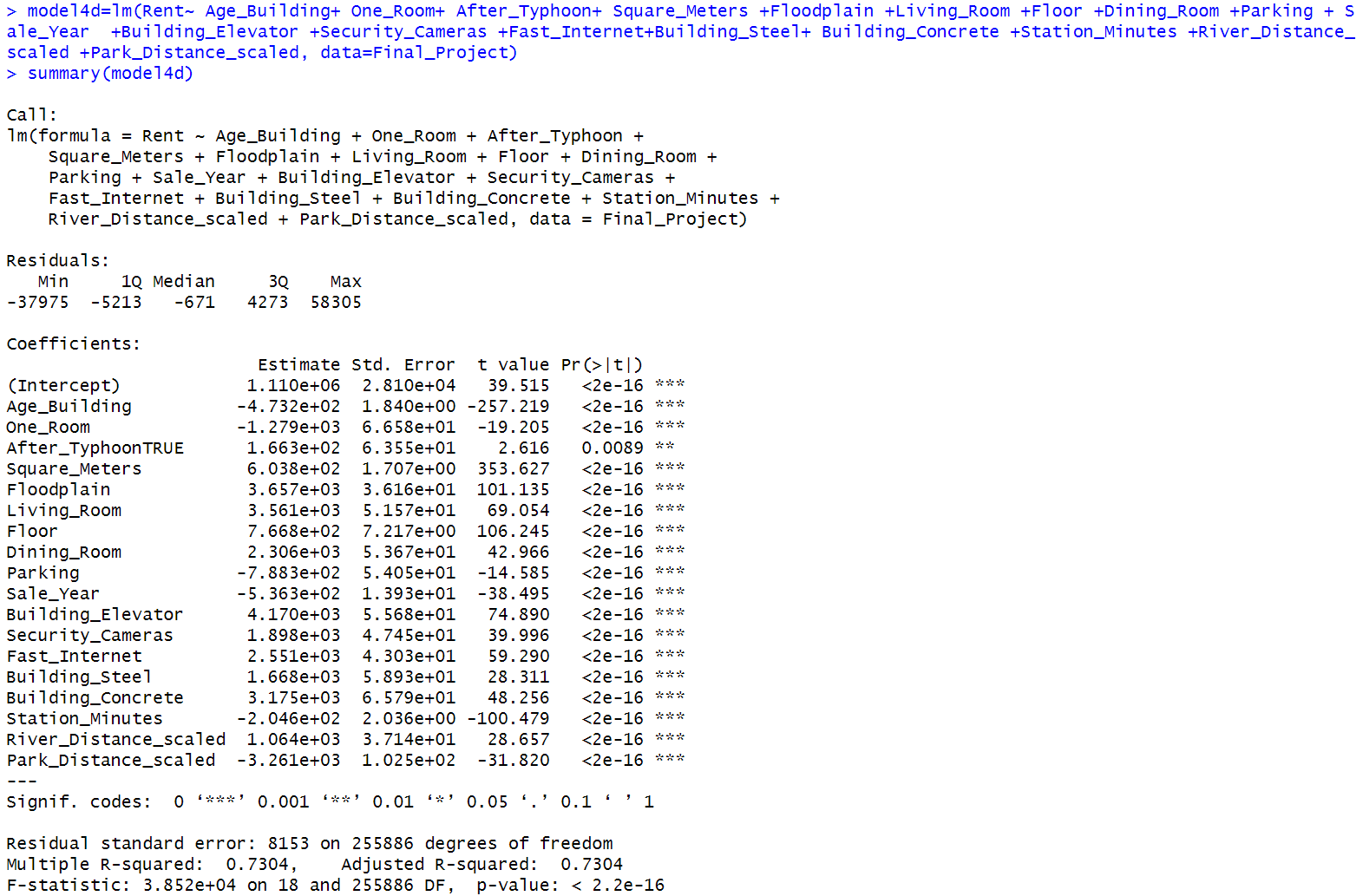
* I then did another chow test to check whether the combination of both variables makes a better model, and as expected it does create a better fit. I will use the unrestricted model (Model2d)



* New model 2d: 
* By adding the “Age\_Building” to the model1, keeping the “Year\_Built” it gives me NA as estimate in the latter variable. This is because of the perfect multicollinearity between the 2 variables meaning one variable can be expressed as a linear combination of the others. In this case, "Year\_Built" and "Age\_Building" are directly related, as "Age\_Building" is derived from "Year\_Built.".

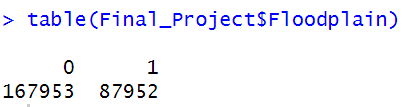


Considering such, I will drop the “Year\_Built” variable and keep the Age of the building instead, since it is easier to interpret (model4d).



This **model 4d** looks a better fit considering that the factors are all statistically and economically significant also in their magnitude. However, the adjusted r-squared and Standard Errors are not much different, which is not affecting much of how much the data explains the model. By adding more variables I am reducing the risk of omitted variable biases but at the same time I am increasing the risk of multicollinearity.

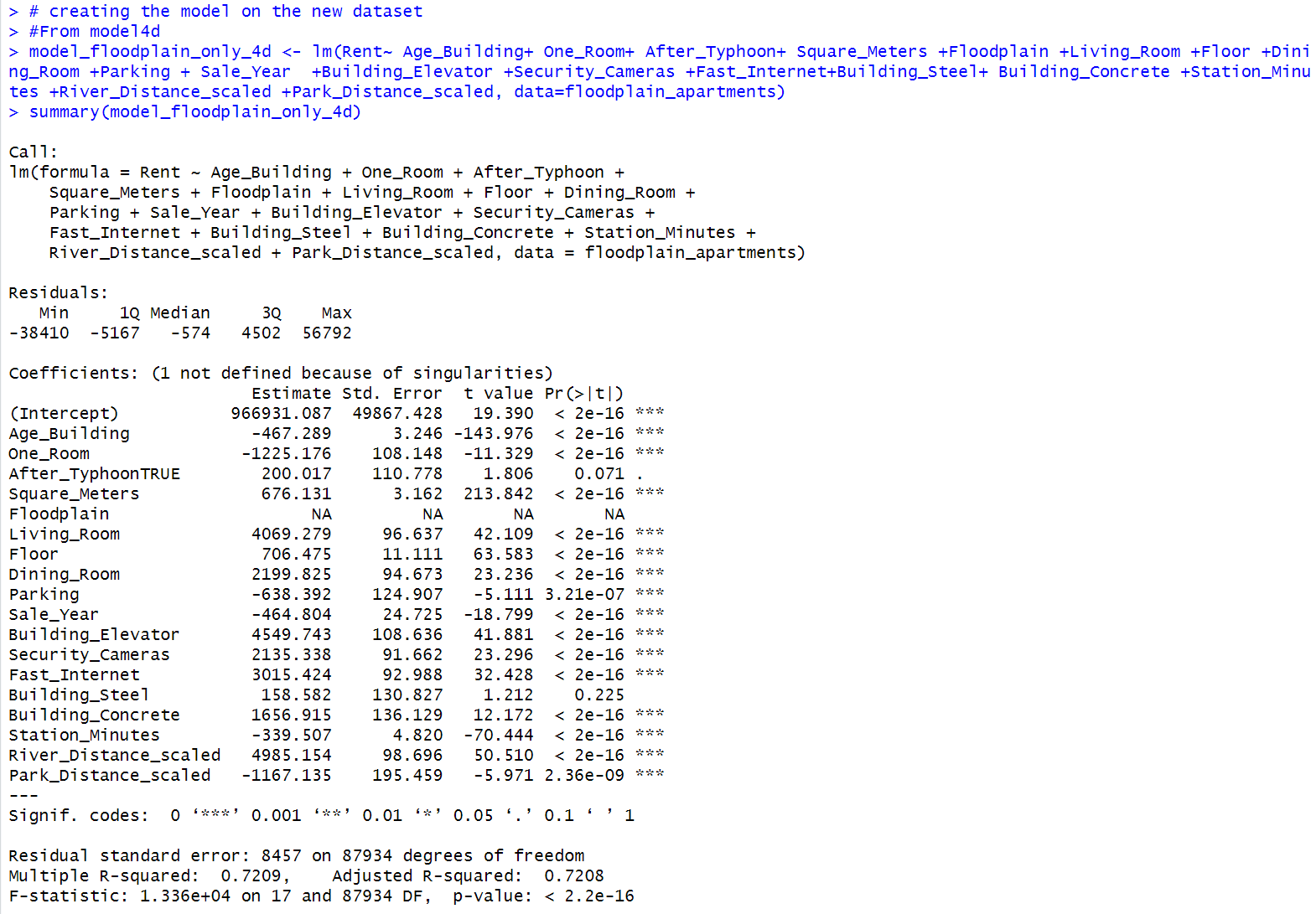
1. Are all apartments in the dataset located within a floodplain? Can you estimate the impact of being in a floodplain on apartment rents if you limited your sample to only floodplain apartments (N=91,639)? **Hint:** I would suggest looking back at the literature review section when answering this question. Based on your answer to this question, will you make any modifications to your model?



**Considering the outliers I removed from the sample in question n°2, my N is slightly lower than 91,639.**

**Looking at the data available, only the 34,37% (N=87952) approximately is located within a floodplain.**

* **BASED ON MODEL 4d**

****

**Considering a dataset with only buildings in a floodplain, there would not be any factor of comparison and we’d get NA since it could not be calculated.**

**We can’t estimate the impact of being in a floodplain on apartment rents if we limited your sample to only floodplain apartments because there is nothing to compare it with. You could estimate it if you add another dummy such as the typhoon month.**

**Considering the literature review apartment rents and housing prices in floodplains do not usually decrease unless there has been a recent natural disaster. But having the dataset that only includes floodplain apartments, may not accurately reflect the overall realistic situation and could not investigate the difference between the treated and the control group.**

**To prevent such, I will consider the model with both the apartments in floodplain and the one that aren’t.**

1. Do you think any of your coefficients are biased due to a missing independent variable? How might this affect your valuation of ? Are you able to fix any of this bias with the information provided in the dataset? Show the output from this updated regression below.

Since I omitted some variables my model could be biased by so. Generally all coefficients are somewhat biased even if the independent variable is uncorrelated with he omitted ones (smearing bias). This would influence the coefficient estimate of and the one of all the other variables.

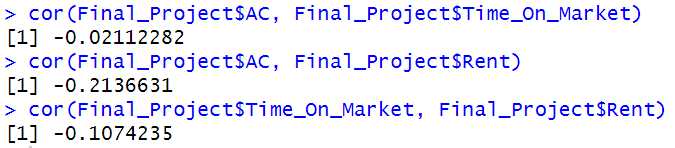
This could be fixed by adding relevant independent variables to increase sampling variance.

Considering my model 4d, with multiple omitted variables, I do not think that any of there would be much difference by adding new variables since I have already analysed carefully why I dropped certain variables. However, I will consider adding

AC and Time on Market .

The potential bias introduced by omitting these variables may affect the valuation of the coefficient for "Floodplain" (β̂\_Floodplain). If AC and Time\_On\_Market are correlated with "Floodplain" and "Rent," omitting them from the model might lead to an omitted variable bias, causing the coefficient for "Floodplain" to be biased and potentially distorted.

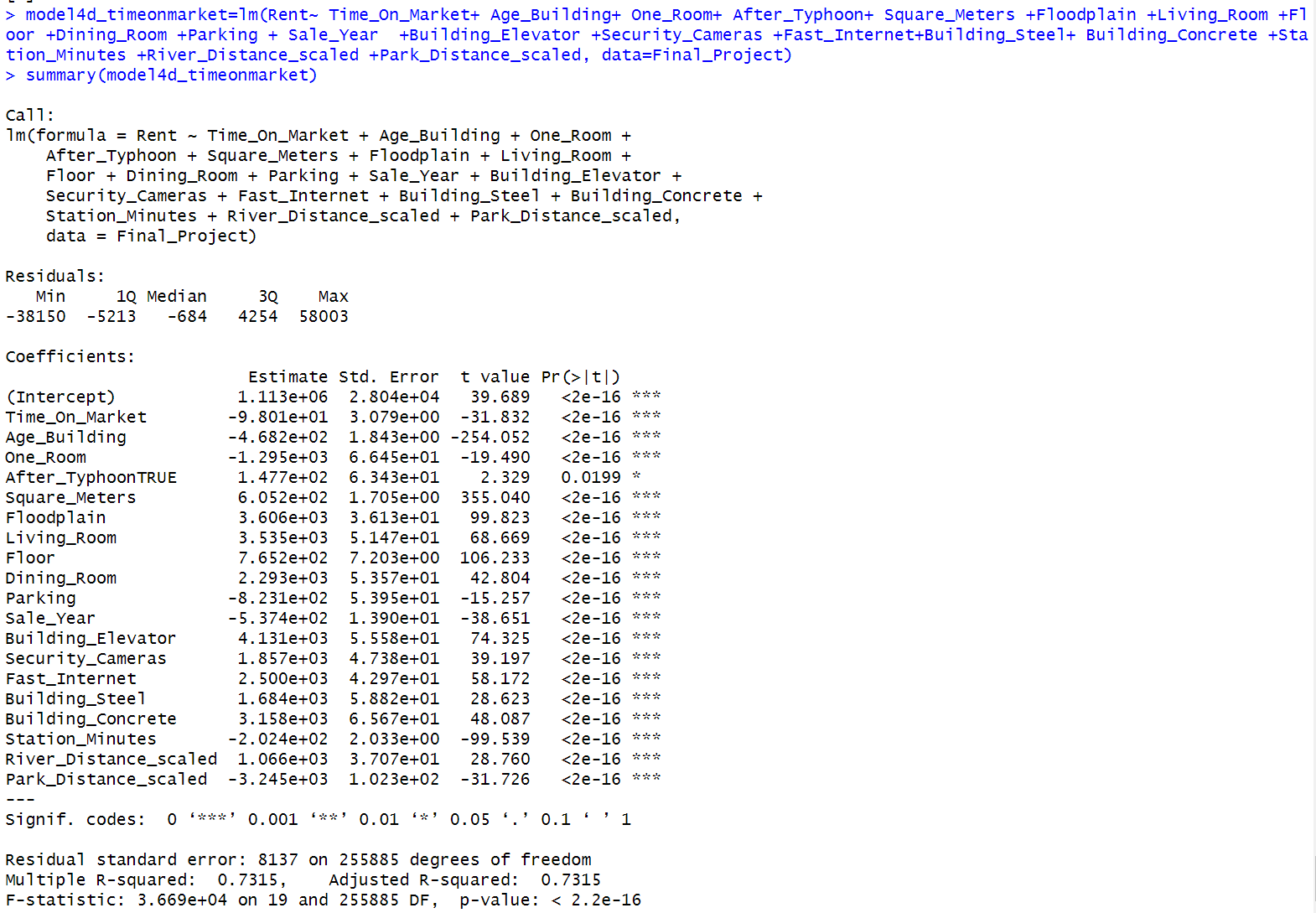
First I check the correlation between the 2 and between each variable and Rent(dependente variable), to prevent collinearity and multicollinearity.



The correlation coefficients between Rent and both AC and Time\_On\_Market are relatively small (-0.214 and -0.107, respectively). Weak correlations suggest that there might not be strong linear relationships between the variables. In such cases, adding both AC and Time\_On\_Market as independent variables may not lead to significant multicollinearity issues, and could potentially be useful for better understanding the impact on Rent.

Looking at the significance of adding them separately, they are both significant(at 5% AC, and 0.1% Time on Market):

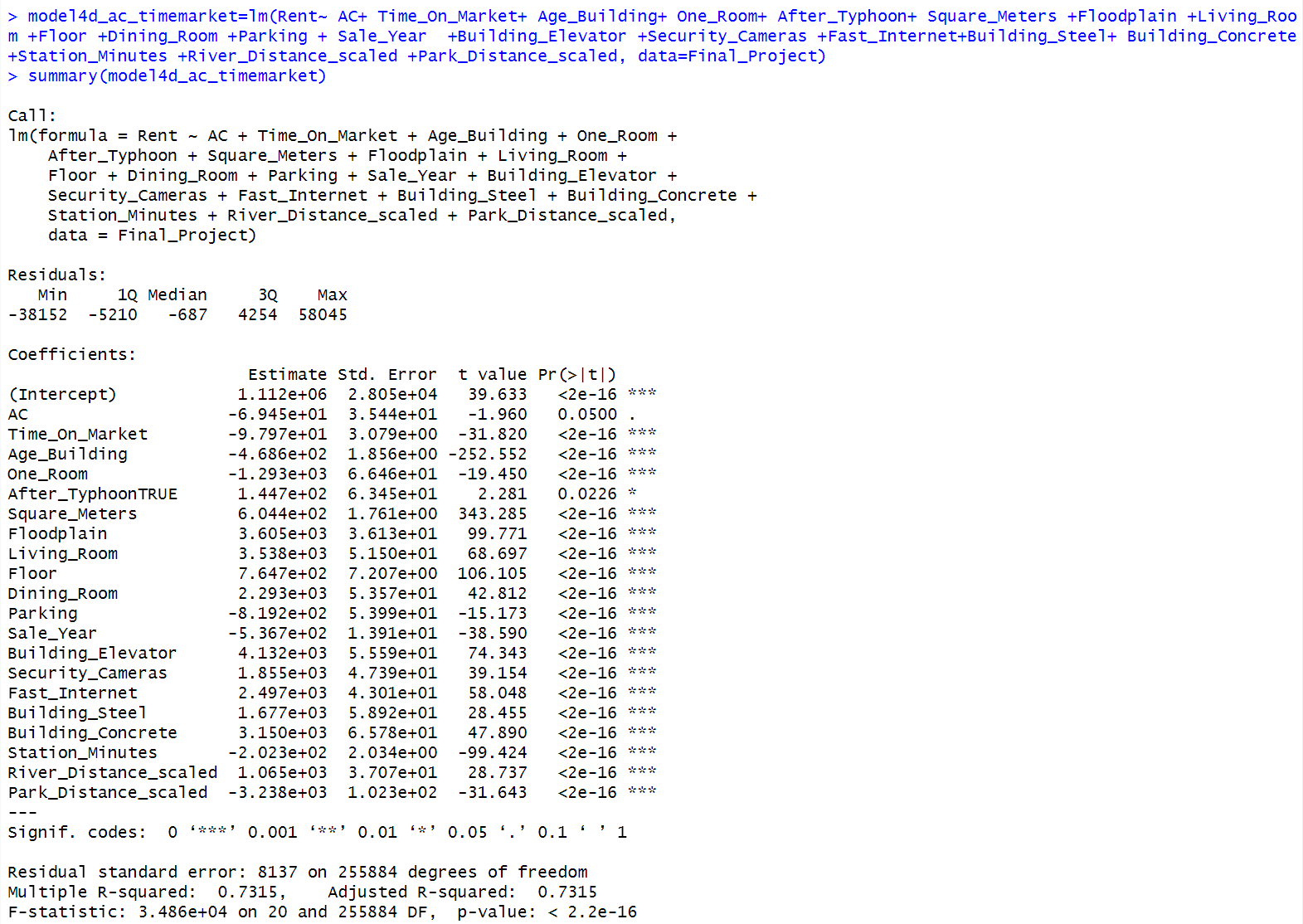
-Time on Market



-AC

A screenshot of a computer

Description automatically generated

-Looking at the impact of adding both the values together I didn’t manage to fix biases. The coefficients and standard errors of the existing variables remained largely consistent , indicating that the addition of these two variables did not have a substantial impact on the estimation of other coefficients. Also, the coefficients for "AC" and "Time\_On\_Market" are both small and not statistically significant at conventional significance levels (p-values of 0.0500 and 0.0089, respectively). This indicates that these variables might not have a strong impact on Rent or do not provide additional explanatory power. Adding more variables to a model does not automatically reduce bias. In fact, introducing irrelevant or weakly related variables can potentially introduce more noise to the model and reduce its efficiency. Additionally, multicollinearity, which can be a result of including irrelevant variables, can lead to inflated standard errors and less reliable coefficient estimates. Comparing this model with model4d, they have similar R-squared values (this model=0.7315 and model4d=0.7304). So adding these variables do not contribute substantially to explaining the variation in Rent. Since adding "AC" and "Time\_On\_Market" did not lead to a significant improvement in the model's fit,these variables as stated previously might be irrelevant or redundant, given that the model 4d already capture much of the information they offer.

1. Check various functional form modifications. Do any functional forms work better than others? How do you know?

Here are some options to consider, though you should think about other functional form decisions as well:

1. Logging your independent variables or the dependent variable
2. Adding quadratic terms as Grether and Mieszkowski did
3. Combining variables by forming “slope dummies” or interaction terms

**I will only consider changing functional forms for continuous variables and not categorical.**

**Looking at the literature, as Raymond Palmquist said: age and price are not linearly related with price.**

* Firstly, I plot the data to check their distribution, which can help deciding which functional form to use.
* **Decision making for each variable:**

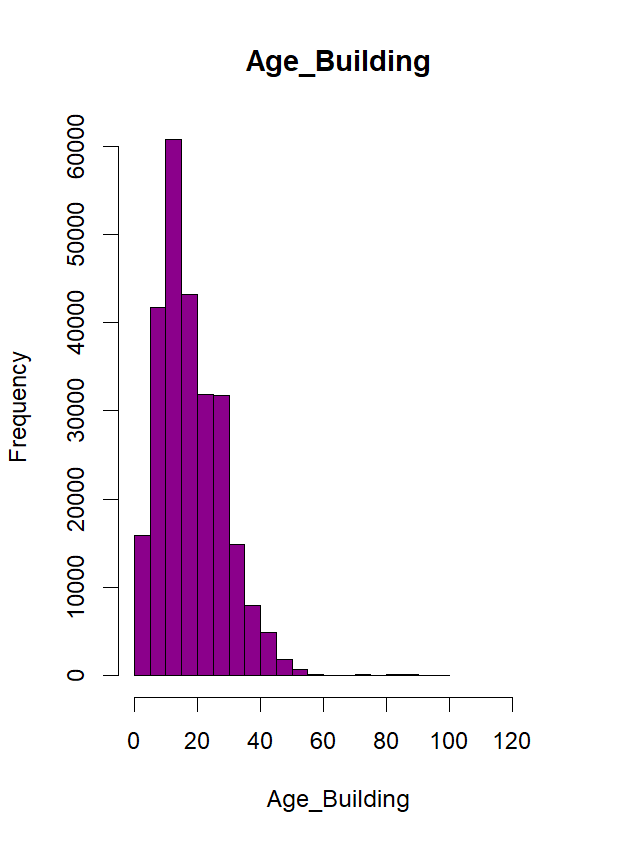
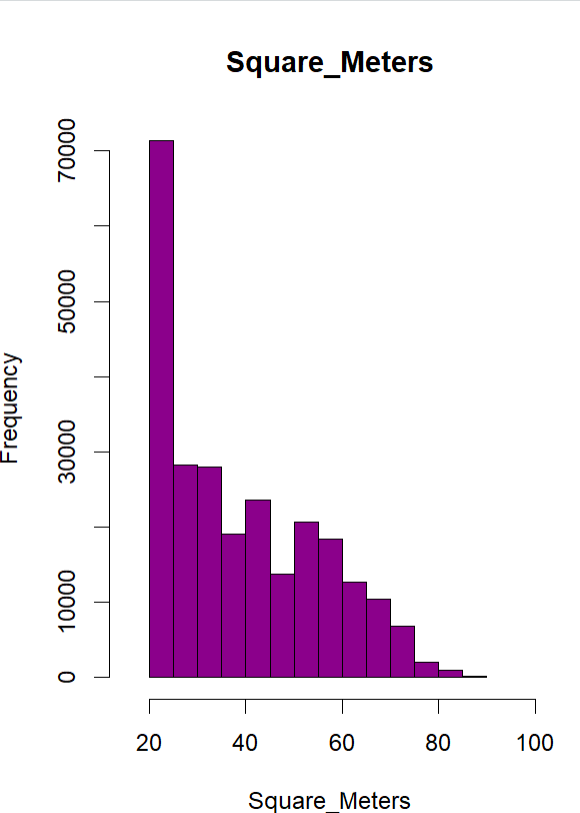
**-Age of the building: It is expected to have a non-linear relationship with rent. As buildings age, their impact on rent may decrease, leading to diminishing returns. In this case, adding a quadratic term may be appropriate to capture the diminishing effect.**

**-Square Meters: The relationship between square meters and rent is often assumed to be linear in many cases. By looking at the graph, it seems that the impact of square meters on rent diminishes as the property size increases instead. I will do a logarithmic transformation.**

**-Station Minutes: I will not change functional form.**

**-River Distance scaled: I will log it, since I suppose that there is a non-linear relationship between rent and river distance. Also, it is what is most commonly used in the literature**

**-Park Distance scaled: I will log it, since I suppose that there is a non-linear relationship between rent and Park distance.**

  A graph of a number of people

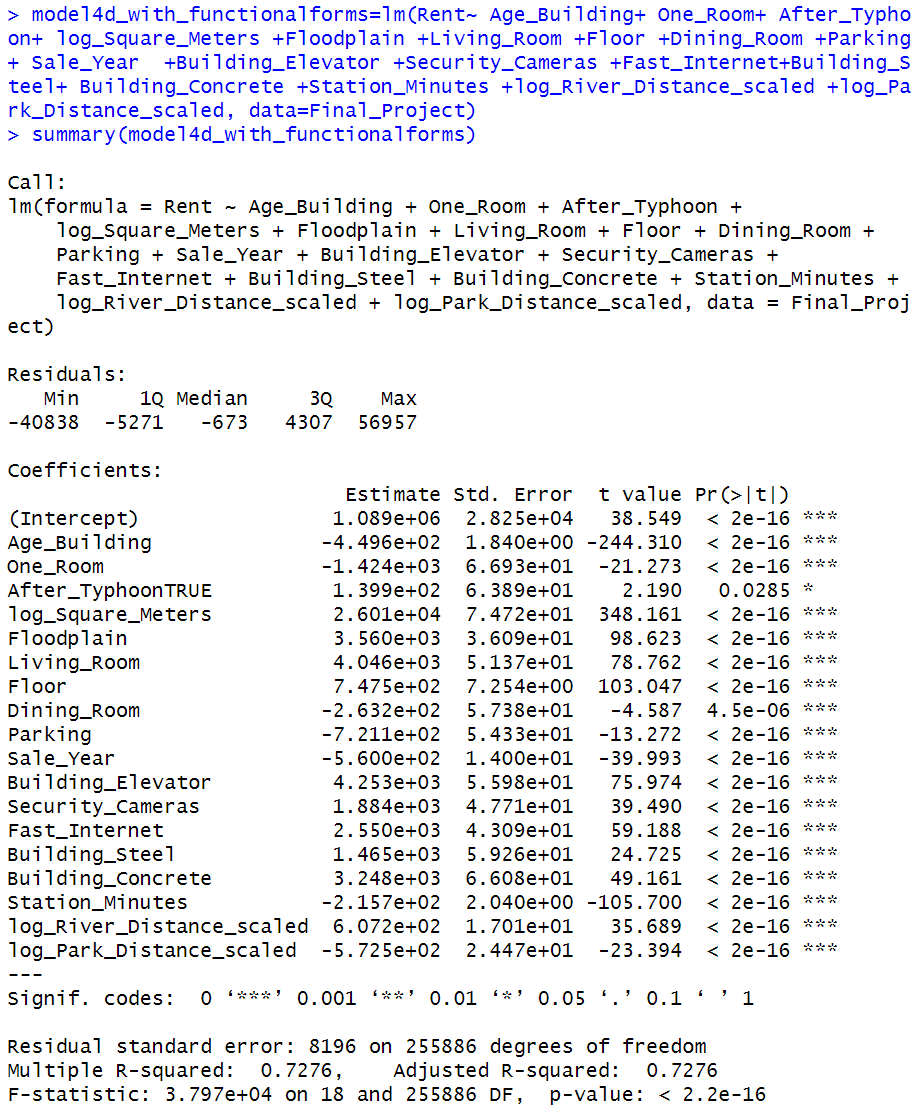
Description automatically generated A graph of a number of numbers

Description automatically generated A graph of a number of numbers

Description automatically generated

* The reason to log is If I think there is a non-linear relationship between rent and independent Variables, or between the variables. This transformation can help to stabilize variance, linearize relationships, and handle data that exhibit exponential growth or decay.

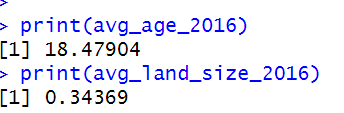
New model with the following changes: After\_Typhoon is now positive,which means that the rent after a typhoon increased by 139,9 yen, keeping all the other variable constants, on average.



* Interaction Terms:

I will interact the dummy month/year in which there was the typhoon (After\_Typhoon) with a variable indicating if the property was located in a floodplain 🡪 B(floodplain\*After\_Typhoon).

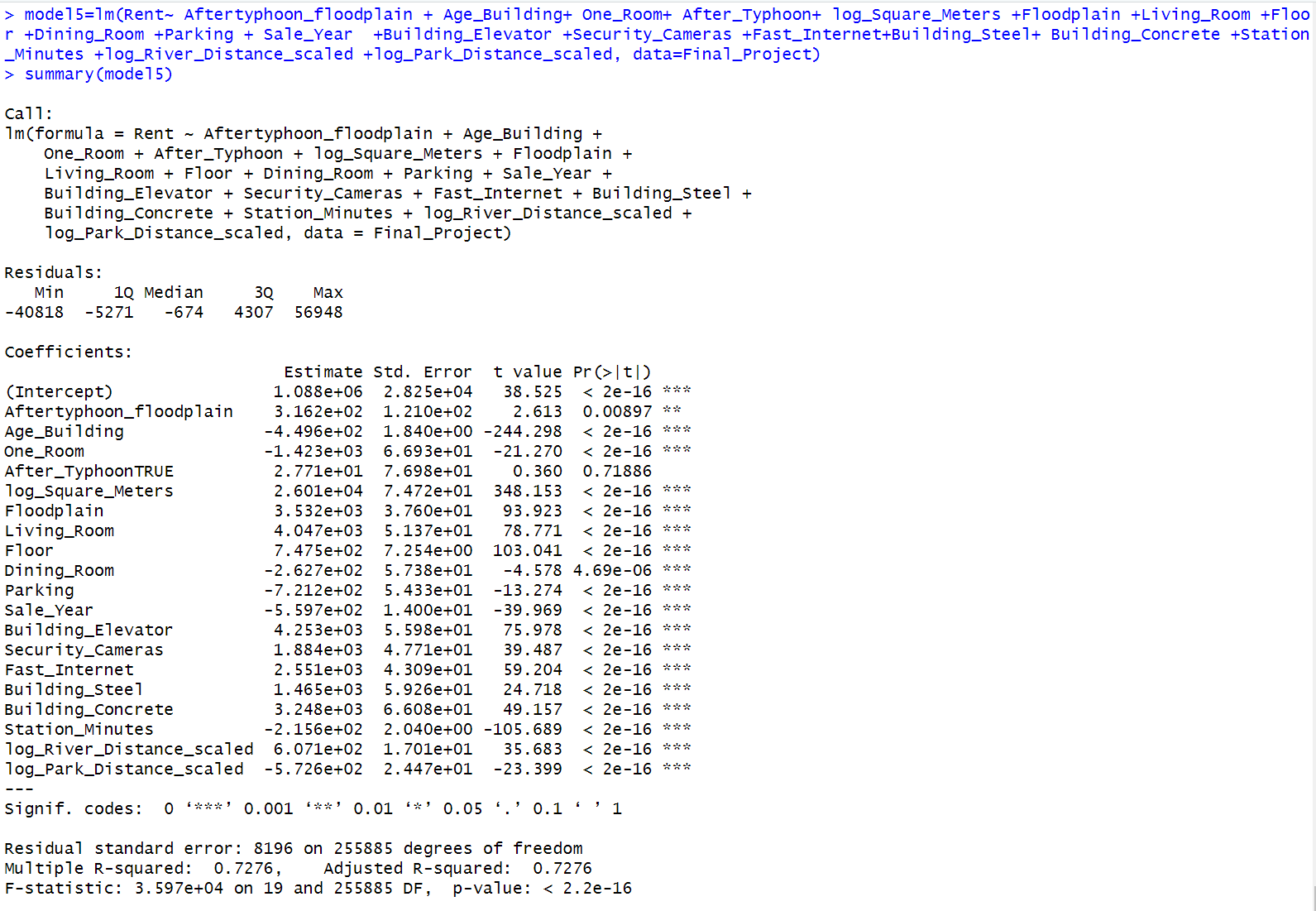
-Then I Checked the sample of houses sampled in 2018 differ in terms of their characteristics than houses sampled in 2016: As a matter of fact there isn’t, which is what we wanted to better analyse and have significant results..



A close-up of a number

Description automatically generated

The new model result(model 5):



No meaningful changes seen looking at adj.R squared and Standard Errors. After Typhoon becomes not statistically relevant.

Interpretation of the result of floodplain\*After\_Typhoon:

A positive coefficient (316.2) for the interaction term indicates that the effect that an apartment in a **Floodplain** has on **Rent** is stronger when **After\_Typhoon** is true (after a typhoon) compared to when **After\_Typhoon** is false. In other words, the presence of a floodplain has a larger impact on rental prices after a typhoon. It aligns with what the literature sais.

(The coefficient for the interaction term **Aftertyphoon\_floodplain** it also has a standard error of 121.0. The t-value for this coefficient is 2.613, and the p-value is 0.00897, which indicates that the interaction term is statistically significant at the 5% level.)

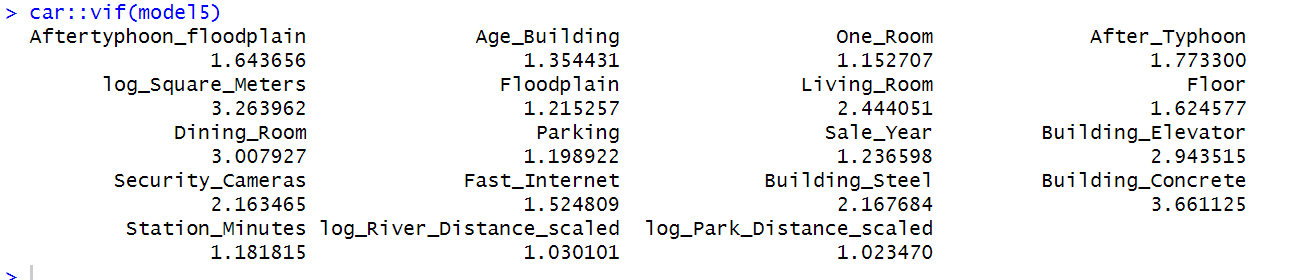
1. Test for multicollinearity. How might this problem affect your conclusions? Which variables may be highly collinear with each other?

**To check for Multicollinearity I will use the VIF SCORE and Correlation metrics:**

**VIF>10 🡪 Multicollinearity**

**5<VIF <10 🡪SOME mult.**

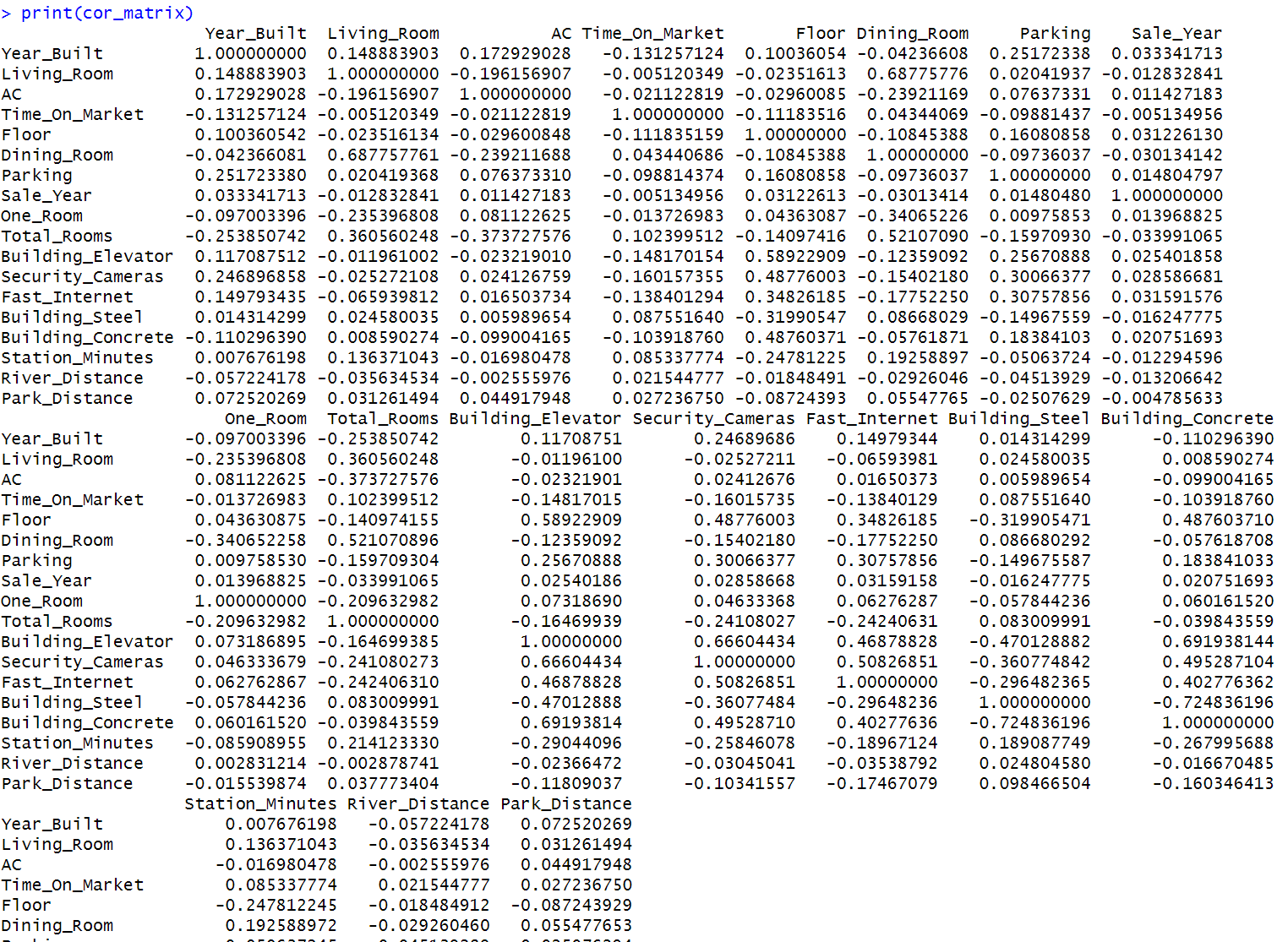
**<5 🡪Low mult.**

****

**Considering the results, the model does not seem to have multicollinearity.**

**This is due to the fact that independent variables are not correlated between each other.**

**In this case there is no sign of collinearity considering a VIF<5.**

**Looking at the correlation matrix:** ****A group of numbers on a white background

Description automatically generated

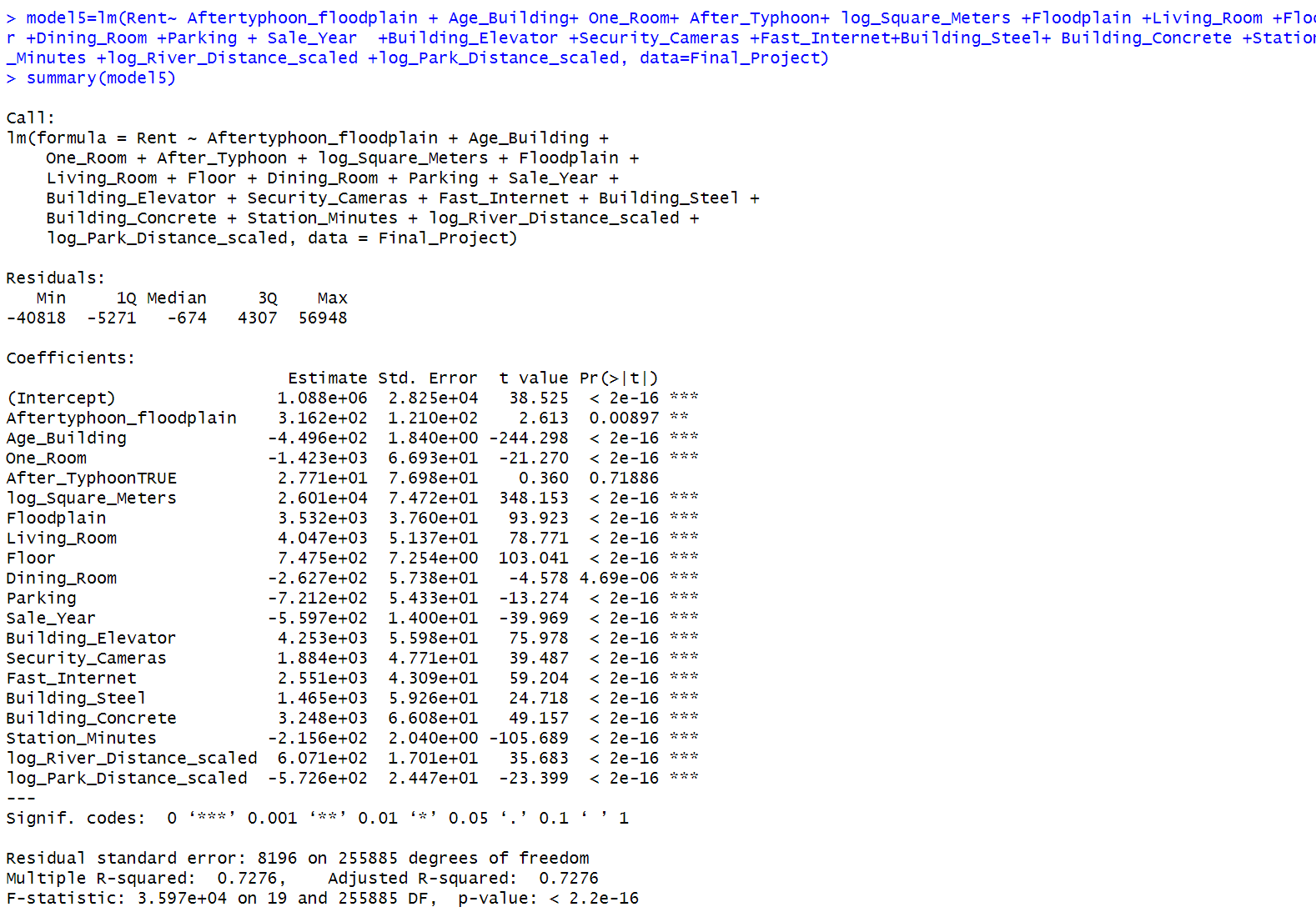
**The variable that show probability of collinearity are:**

* **Building concrete and building elevator (0,69🡪 closer to 1);**
* **Building steel and building concrete=-0.71 (negative moderate correlation, but closer to 1);**
* **Dining Room and living room with 0.68 of positive correlation;**
* **Buinding elevator and floor (0.58);**
* **Building elevator and security cameras (0.66).**

**Considering that After\_Typhoon is not significant, to adjust it I removed Sale Year since they are moderately correlated.**

****

**Before:**

****

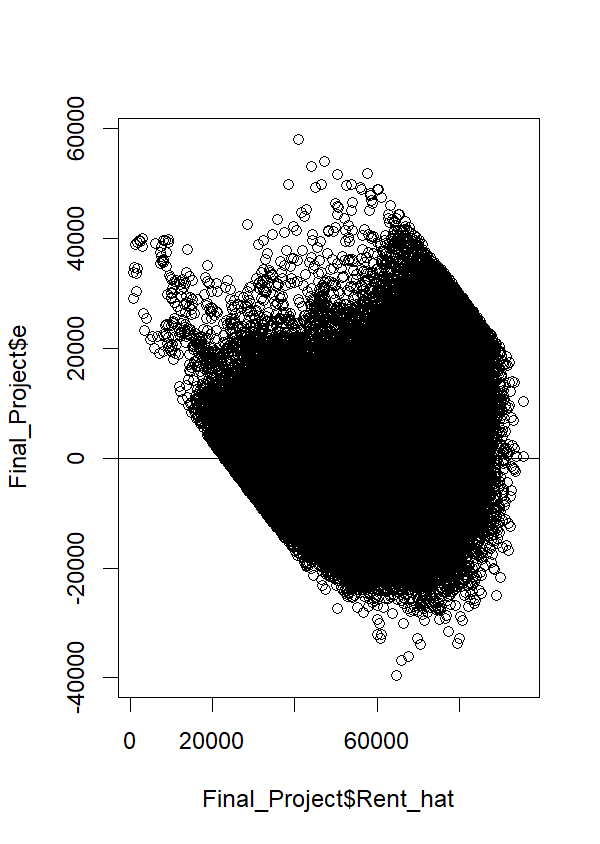
**After dropping it, not much influence either in the Adjuster R squared nor the Standard Errors and p-values(model 5 test):**

**A screenshot of a computer

Description automatically generated**

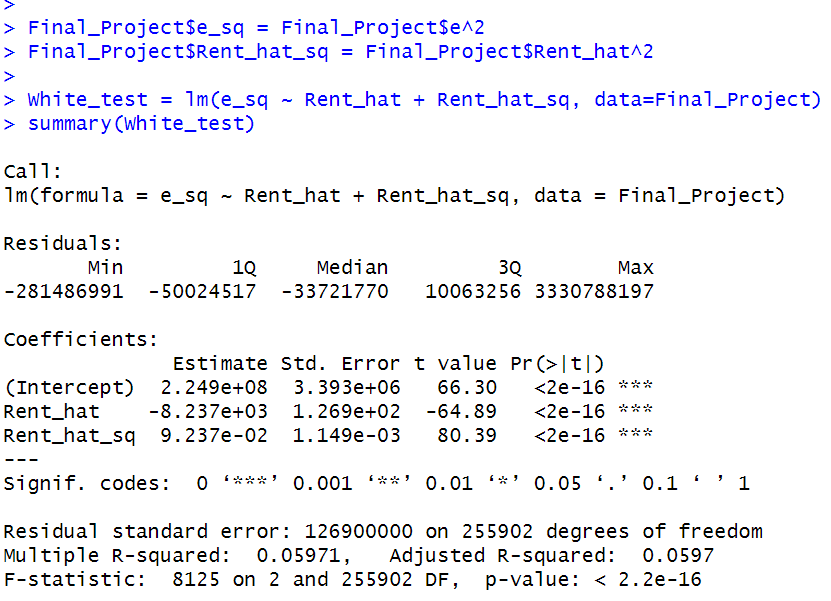
1. Test for heteroskedasticity and serial correlation. How might these problems affect your conclusions? What might be causing these problems? If you detect either problem, how will you fix it? Please show what happens to your model when you apply these corrections.

* **Heteroskedasticity:**

****

**There is evidence of heteroskedasticity looking at the plot.**

**By running the white test to detect heteroskedasticity, more specifically we can see that there actually is considering the very low p-values. It is highly significant, we reject the null Hypothesis (reject homoskedasticity) and accept Ha (accept heteroskedasticity). The dispersion of the error term changes with the estimated rent\_hat.**

****

* **Detect Serial Correlation:**

**-Calculate the Durbin Watson test**

A computer code with numbers and letters

Description automatically generated

**The autocorrelation value of 0.3625842 indicates that there is some positive correlation but moderate between the residuals of the regression model. Consequently, there might be additional variables or factors that should be included in the model.**

**The Durbin-Watson test statistic is 1.2748, which is less than 2. This suggests that there might be positive autocorrelation in the residuals.**

**A value close to 2 in the Durbin-Watson test suggests that there is likely no serial correlation. I would interpret the result as been closer to 2 or there might be just some.**

**The p-value of 0 instead indicates that the observed Durbin-Watson statistic is statistically significant, which means that**

**we have evidence to reject the null hypothesis of no autocorrelation in favor of the alternative hypothesis of positive autocorrelation.**

* **Fix both serial correlation and heteroskedasticity**

**We use the heteroskedastic-corrected SE, which is the most commonly used. A screenshot of a computer program

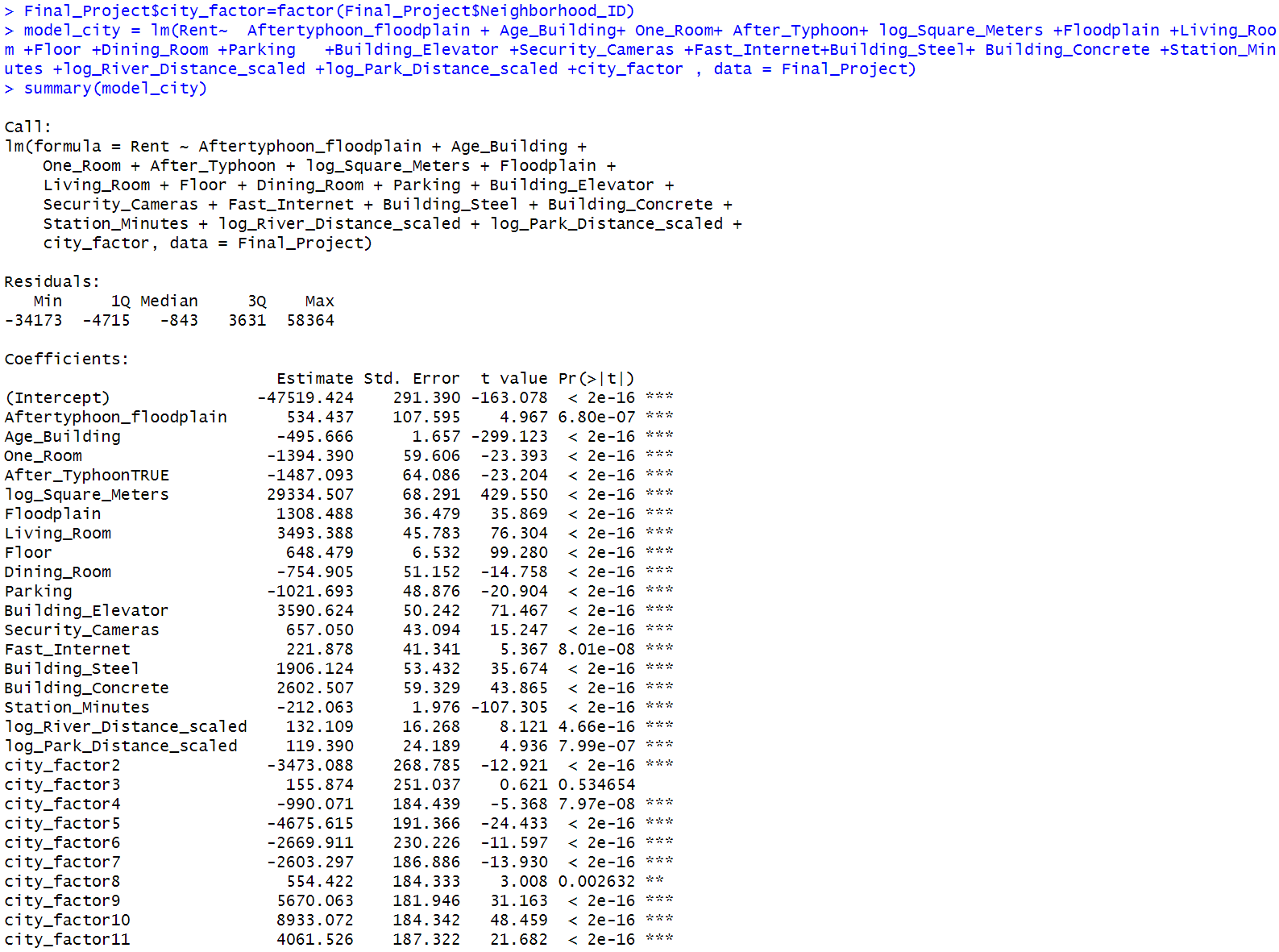
Description automatically generated**

**(We are not using weighted least square since the sample is big and does not make sense to use it here. It would create biases.)**

**Looking at the results all the Standard Errors are the same which is not what we would expect by running this model, whilst the coefficients stay the same (as expected). It means that either there was no serial correlation or that it did not fix it.**

1. Included in the R Datafile is a variable titled , which is an index that identifies which neighborhood the apartment is located in. For example, you can think of as the following: = 1 for apartments near Sannomiya station, = 2 for apartments near Rokko station, etc. Using this variable, it is possible to include a dummy for each neighborhood in the regression model. Should you include neighborhood dummies in your regression, why or why not? What do these neighborhood dummies control for that isn’t already included in the model?

To answer the following

 A vertical line of numbers

Description automatically generated A white background with numbers and letters

Description automatically generated

**Firstly, by looking at the data, I can see that some city factors are missing since they are already missing in neighborhood data itself.**

**With this approach we are trying to analyse separately each factor. Basically, including the city dummies we can actually see the coefficients for each neighborhood variable to be able to analyse how on average rent varies from one neighborhood to the next. (With the within estimator this is not possible).**

**The results are the same because when including the Neighborhood dummies(city factor) in the rent regression, we are holding the effect of the dummy constant( the impact of being in different neighborhood).**

**Each city\_factor represents a specific city, and the corresponding coefficient estimates tell us how the rental rate varies for each city compared to the reference city 1 (Our base category).**

**For example, let's take "city\_factor2" as an example:**

* **city\_factor2 coefficient: -3473.088**
* **p-value: < 2e-16 (extremely low)**

**The negative coefficient (-3473.088) indicates that the rental rate in "city\_factor2" is lower than the reference city (base category). The p-value being extremely low suggests that this difference is statistically significant.**

**Many cities have positive or negative rental rate compared to city\_factor1(the omitted), with a certain amount in yen higher or lower rental prices than city1, holding all other variables constant.**

**For example city factor 59 ha a price -18543.107 yen lower than city 1, holding all other variables constant. I would interpret this neighborhood as being maybe in a better neighborhood, closer to the station. Looking at the data I can see that there are way less data of apartments in neighborhood 59 than neighborhood 1, so this could be impacting some biases.**

**Why should I include these factors to the model?** **N**

**Including Neighborhood dummies can help control for unobservable factors specific to each neighborhood that are not explicitly included in the model. These unobservable factors could be neighborhood characteristics, amenities, local regulations, or any other unique attributes that influence rental rates but are not captured by the variables in your current model.**

**It can also improve the interpretability of this model. By estimating separate coefficients for each neighborhood, you can understand the specific effect of living in a particular neighborhood on the rental rate, holding all other variables constant.**

**It can help capture the heterogeneity of difference in rental prices between neighborhood. Of course including many dummies could also lead to collinearity issues since we have a big sample data.**

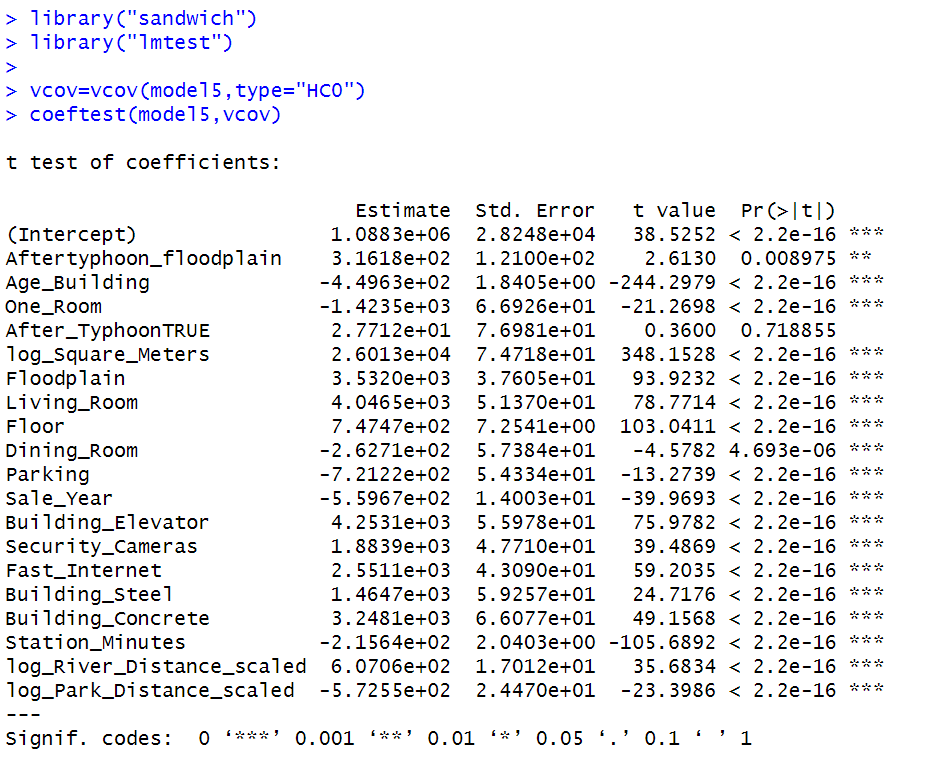
1. At the end of this course, we discuss limited dependent variable models. Should we use the methods outlined in this session to model rental price? Why or why not?

**We cannot use it, since our dependent variable rent is not binary.**

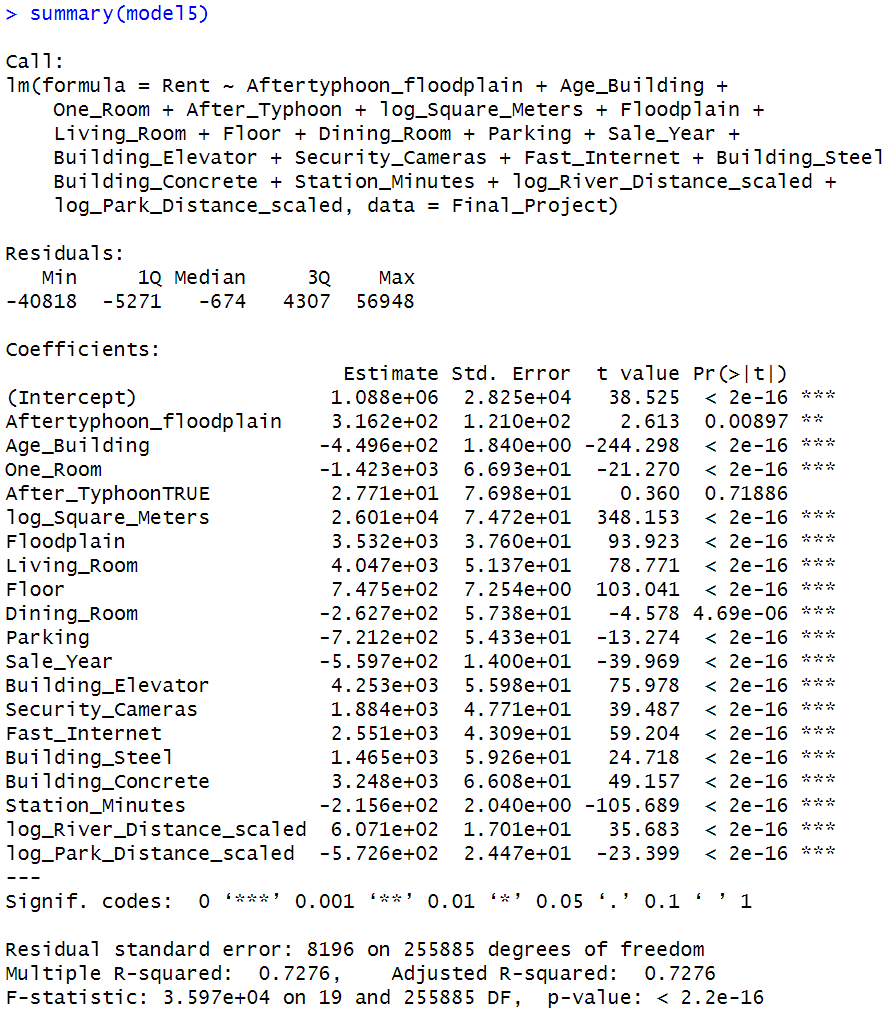
1. Create a new regression model which fixes some if not all the problems you discovered in questions 2 and 4 - 12. How has this new model improved upon your old model (i.e. are more independent variables significant, do all the coefficient magnitudes now make sense, etc.)?

**As done in Question 9, I have already done some changes. The entire process till the white test, showed us a better fit , fixing problems such as heteroskedasticity.**

**Since the fix heteroskedasticity with the het.corrected SE did not work I will perform it considering the model5:**

****

**In this case, also in this case it did not provided any adjustment. This is again not expected. Maybe it’s because there is a weak heteroskedasticity, or I did some computational errors.**

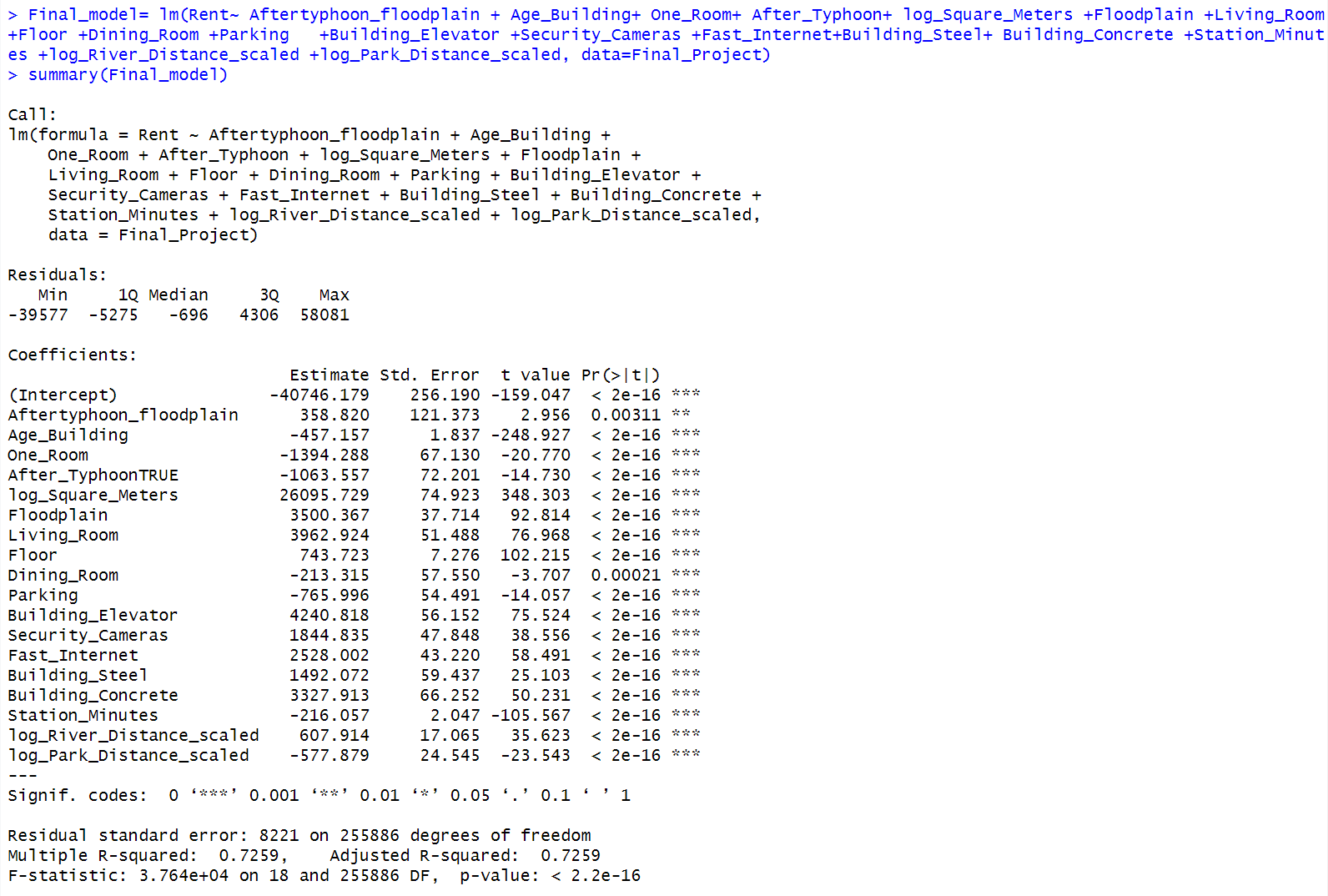
****

1. Can you use an F-test to determine if your new model provides a better statistical fit of rental prices than your old model (i.e. from question #3)? Why or why not? If an F-test is not appropriate here, then compare across the two models using a different statistical measure. If an F-test is appropriate, compare across the two models using this statistical measure. What can you conclude from this analysis?

**Running an F- test would not be appropriate since the 2 models are very different. We dropped some variables.**

**We can use R-squared though, or other measures like looking at the coefficients.**

1. Once you are satisfied with your final model, please include your regression output from your new model below. Does your valuation of align with what was found in the literature? Why or why not?



**Regarding the coefficient of interest, which is the impact of living in a floodplain on apartment rental prices (β̂\_Floodplain), it is estimated to be 358.820. The coefficient is statistically significant at the 0.05 level (p-value = 0.00311), indicating that living in a floodplain has a significant effect on apartment rents.**

**The estimated coefficient of 358.820 suggests that, on average, apartments located in floodplains have monthly rents that are 358.820 Japanese Yen higher than those not located in floodplains, holding all other variables in the model constant.**

**Comparing this estimate with the literature findings, it appears to align with the general notion that living in riskier areas (such as floodplains) would result in lower apartment rents. Although the literature indicates that the effect might be more pronounced after a natural disaster, the estimated coefficient in this model (358.820) still shows that there is a statistically significant impact on rents due to living in a floodplain.**

The coefficient for After\_Typhoon in the model is estimated to be -1063.557. The negative coefficient suggests that, on average, apartments leased after the typhoon have monthly rents that are 1063.557 Japanese Yen lower than those leased before the typhoon, holding all other variables in the model constant.

Looking at the logged distance variables:

1. log\_River\_Distance\_scaled: This variable represents the log-transformed and scaled distance from the apartment to the nearest river. The positive coefficient of 607.914 indicates that as the distance from the river increases by 1%, the apartment's rent tends to increase by 607 yen on average. In other words, apartments that are farther away from the river are associated with higher rents, all other variables being held constant. It resonates with my expectations.
2. log\_Park\_Distance\_scaled: This variable represents the log-transformed and scaled distance from the apartment to the nearest park. The negative coefficient of -577.879 indicates that as the distance from the park increases by 1%, the apartment's rent tends to decrease by -577 yen on average. In other words, apartments that are farther away from parks are associated with lower rents, all other variables being held constant. It resonates with my expectations.

**Works Cited**

Atreya, A., Ferreira, S., & Kriesel, W. (2013). Forgetting the flood? An analysis of the flood risk discount over time. *Land Economics*, *89*(4), 577-596.

Bin, O., & Landry, C. E. (2013). Changes in implicit flood risk premiums: Empirical evidence from the housing market. *Journal of Environmental Economics and management*, *65*(3), 361-376.